Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No 3 special issue (2020) 1-9
ISSN 2295-3159

The Promise and Pitfalls of Self-report: Development, research design and analysis issues, and multiple methods.

Luke K. Fryer[a] & Daniel L. Dinsmore[b]

[a] The University of Hong Kong, Hong Kong

[b] University of North Florida, USA

Abstract

As a prelude to this special issue on the promise and pitfalls of self-report, this article addresses three issues critical to its current and future use. The development of self-report is framed in Vertical (improvement) and Horizontal (diversification) terms, making clear the role of both paths for continued innovation. The ongoing centrality of research design and analysis in ensuring that self-reported data is employed effectively is reviewed. Finally, the synergistic use of multiple methods is discussed. This article concludes with an overview of the SI's contributions and a summary of the SI's answers to its three central questions: a) In what ways do self-report instruments reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use? b) How does the use of self-report constrain the analytical choices made with that self-report data? c) How do the interpretations of self-report data influence interpretations of study findings?

Keywords: self-report, multiple methods, vertical and horizonal development, research design and analyses

Corresponding author: Dr. Luke Fryer, University of Hong Kong, fryer@hku.hk. DOI: https://doi.org/10.14786/flr.v8i3.623

1. This SI's Mission

While self-report measures are ubiquitous across and often central to educational research, they are also often denigrated for a range of reasons. For instance, the reliability of the measures and the validity of the resultant score interpretations are often called into question (e.g., Veenman et al., 2006). This has led to calls for moratoria on the use of self-report in some corners. However, rather than discarding or ignoring data generated from self-report measures of cognitive processing, motivation, emotions and beliefs, research is needed to determine *when* and *if* self-report measures can contribute to our collective understanding of theory surrounding these constructs. For example, relying solely on self-report to study regulatory processes has contributed little to our understanding of self-regulated learning (Dinsmore et al., 2008), however, in other instances self-report may be the only viable manner in which to unearth covert constructions, such as self-efficacy (e.g., Zimmerman, 2000).

This special issue examines the accuracy of interpretations and conclusions drawn from self-reports regarding individuals' metacognitive and cognitive processing, affect and beliefs, and the analytic choices made. These questions are addressed by an international group of experts examining these constructs from different theoretical and analytical perspectives. The current Special Issue as whole brings three issues that are often noted, but rarely specifically discussed into focus: lateral versus horizontal development of measurement approaches, the critical role of research design and analyses, and the complex role of utilizing multiple measurement methods.

2. Lateral and Vertical Innovation: Both are critical

The first topic to be addressed in this Special Issue is how self-report approaches are advancing both laterally (i.e., improving current methods) and horizontally (i.e., developing new methods). As an analogy, a vibrant city in the late 19th and early 20th century was bustling with horses, horse-drawn trams, and cars that coexisted along the city thoroughfares. Figuring out the best way to get across the city is dependent upon many factors – such as the persons wealth or when they are trying to make their transit. Similarly, the research literature is replete with different vehicles to transport oneself from Point A to Point B – namely how to measure the processes described in this special issue to better understand theory, and ultimately, student learning. As with the city, the best mode to measure these constructs depends on many factors. However, unlike the advances in transportation technology, the advances in and pressure to modernize self-report methods has been weak at best.

This advancement of methods (or lack thereof) for measuring latent constructs critical to educational research (self-report included) can be framed by Horizontal versus Vertical conceptions of development. This is a well-established framework for understanding growth (e.g., economic innovation; Bondarev, & Greiner, 2019) and change (e.g., natural selection; Lawrence, 2005) in a broad array of fields. Horizontal growth refers to innovating towards entirely new approaches, while vertical growth refers to refining and enhancing current methods. This framework fits the current era as educational research is flooded with new (Horizontal) means of measuring students' cognitive processing (e.g., eye tracking; Chaulic et al. 2020) and meta-cognitive processing (e.g., trace data; Rogiers, et al., 2020). This Horizontal drive for innovation of measurement continues to push into complex areas such as emotion (facial recognition; Chiu, et al., 2019; Dingle, et al., 2016; skin conductance, Järvenoja, et al. 2018; Lehikoinen et al., 2019) and motivation (neuroscience; Hidi, 2016; Mayer, 2017). The considerable momentum behind this drive for alternatives to self-report measurement arise in large part with a longstanding dissatisfaction with their intra-psychic nature and the general lack of Lateral development in these measures.

Given the attention that the horizontal development of these measures has garnered and the lack of lateral development, this special issue addresses the many areas of lateral development that are possible. In other words, this special issue forges new inroads towards further development of self-report measurements

across a range of processes. Not only do these empirical pieces suggest lateral development is possible, each starts to take us down this journey and provides evidence that these journeys are likely to be fruitful. From empirical multimethod studies (e.g., van Halen et al., 2020; Rogiers, et al. 2020) to the theoretically-rich commentaries (Pekrun, 2020; van Meter, 2020; Winne, 2020;), this Special Issue suggests that self-report measures are a unique, valuable – and therefore irreplaceable – source of information about many critical aspects of the learning processes under study here. Clearly, the conclusions drawn from these analyses show that self-report remains critical in our understanding of learning *and* that educational researchers need to push harder for constant lateral innovation such as these. It safe to say, however, that many of these researchers have not felt this obligation. As noted in Fryer & Nakao's (2020) contribution, the primary self-report instrument for the majority of educational research is a tool invented in the 1920s. On paper or smartphones, a Likert scale is a Likert scale. We can and should be struggling to improve on the tools of our trade – as is already being done widely across the technology industry (e.g., Google; Lawless & Biel, 2020).

Lateral advances in self-report can take many forms, a few of which are presented in this Special Issue. Addressing perceived weaknesses in the format by either adding to it (Durik & Jenkins, 2020) or changing it (Fryer & Nakao, 2020) are both rungs in the ladder up toward vertical innovation. Addressing how self-report tools are used and their results analysed (Moller, et al) is another means of climbing further up those rungs. A scan of leading journal suggests that the latter approach to lateral innovation is expanding (e.g., Gillet, et al., 2019; Yuen, et al., 2019), while the former – i.e., analysis – is almost unknown. Both are necessary if measurement of constructs critical to education (self-report or otherwise) is to continue to improve.

3. Not compounding self-report error: It is just common sense

The limitations sections of educational research articles are replete with apologies. There are three apologies that in our experience vie for most prevalent: a) the self-reported nature of the data, b) the cross-sectional nature of the research design and, hinging on the first two, c) the rigor of the analyses. It is time researchers stopped apologising for the first when it is necessary and useful, and did something about the latter two. The weaknesses of self-report have been well known for some time. As the present Special Issue has confirmed, however, self-report also has its unique strengths – often left unmentioned in critiques of self-report. Self-report is an important part of research in many areas (e.g., motivation), but should not be the *only* measurement tool. The supplement of self-report with other observed measures has the potential to improve our understanding of the complex interrelations between the variables described in this Special Issue and learning. This balanced approach to measurement should bring an end to self-report's inclusion in limitations sections. The unresolved issue for the field is that many researchers continue to compound the weaknesses of self-report with cross-sectional designs and inappropriate analyses. These are at least partially linked, as authors, struggling to get published, are desperate to use popular analytical methods. The best example of this mis-use is structural equation modelling (SEM) with cross-sectional data sets. The often small sample sizes being utilised means that researchers are forced to use mean-based SEM rather than latent-based SEM (for an introductory overview of each and their differences see Kline, 2011). This pairing of common analytical mis-steps, compounds the inherent weaknesses of self-report: like building a tall, narrow building in an earthquake prone area.

These two research design issues amplify the self-report concerns cited previously in separate but connected ways. First, educational research is generally seeking to explain learning or learning related processes; processes which are by their very nature developmental and require longitudinal examinations of that development. Research using snapshots of the learning experience can make meaningful contributions to educational research, but only if the limitations of these static designs are taken seriously and appropriate analyses employed. Ginns et al. (2017) is an example of exactly this kind of theoretically robust, carefully structured cross-sectional research. Second, self-report within educational research is generally employed to measure latent constructs. It therefore makes sense to employ analyses that treat self-report data as though it were representing latent constructs. This seems especially pertinent to survey self-report which generally

measures constructs with multiple items. For fine grained analysis of subtle aspects of the learning process or interventions having nudge like effects (small but meaningful over time), the error imbibed by averaging across multiple survey items is a serious, and too often ignored issue. Novice readers and those skimming through articles, are prone to conflate the often mean-based and latent-based SEM. Path analysis, in addition to its inherent mistreatment of latent variables, prevent fully forward analysis due to a lack of degrees of freedom, resulting in the picking and choosing of regressive connections. This additional author-induced source of error is akin to the file-drawer bias (i.e., you don't get the whole picture). Readers of articles that utilize mean-based SEM are too often presented with a cropped picture, only showing connections which support the researcher's aims. What is commonly referred to as path analyses is just one example of how self-reported data can be mishandled, and lead to exacerbating their inherent weaknesses.

It is important to restate that cross-sectional designs can make a limited contribution to research, but researchers need to acknowledge their limitations and not draw conclusions that their data does not support. Researchers seeking to make a strong contribution to an area of educational research where self-reported measures are an important part of quality research design should strive to employ designs that can capture developmental processes and analyses that recognise latent for what it is: unseen. For a detailed and balanced discussion of this issue, we encourage a careful review of Martin (2011).

4. The promise of multiple methods

As discussed previously, many of the papers in this special issue use multiple measures to present a more complete picture of the complex interrelations between constructs. However, we should be careful to distinguish between the aims of engaging in this process and the analyses of these multiple measures, which has often been referred to as *triangulation* (Godfroid & Spino, 2015). We see three potential paths here: using measurements to identify the same aspect or aspects of a constructs, using measurements to identify complementary aspects of a construct, or both. We offer an analogy here to help the reader better understand these paths.

The first path, identifying the same aspect would be akin to using multiple measurements of sound to identify the pitches (how high or low a note sounds) of the notes in a melody (i.e., the part of a tune you might hum). One might use a well-trained ear and an electronic tuner to do this. If both the listener's ear and the tuner are accurate, they should agree on the pitches – maybe the tune starts out and ends on "middle C". Similarly, when examining one of the constructs in this Special Issue, say metacognition, this first path would be akin to saying that our multiple measurements are indeed measuring the same aspect of metacognition – that they should agree. This approach would often be analysed using a multi-method multi-trait analysis (MTMM; c.f., Campbell & Fiske, 1959). Here, we expect the same techniques used to measure the same construct to "agree" more often than those techniques used to measure different constructs. For example, if retrospective self-report and physiological measurements are used to measure reading comprehension and mathematical achievement, the self-report and physiological measurements of the same construct (e.g., reading comprehension) should correlate more closely than the two self-report measurements of the two different constructs. The latter would be an example of a *methods effect*, while the former would demonstrate that both measurements are measuring the same aspect or construct.

However, the melody of a piece of music is often not the only aspect of a musical composition. In a symphony, for instance, the melody is often accompanied by other lines of music as well (which would be harder for the novice to hum). Thus, it might be necessary to use different techniques to identify the sounds present. While a well-trained ear would be able to pick out the chords composed by these multiple musical lines, a simple tuner would not. This issue gets even more complex when thinking about a Bach fugue for example, that layers multiple melodies and countermelodies to create a rich tapestry of sound (c.f., J. S. Bach's Toccata and Fugue in D minor, BWV 565). This more complex conceptualization describes the second path here – are we using multiple measurements to better describe the rich symphony of a process at play? In other words, are there multiple aspects of a particular construct that some measurements are better at tapping than

others? For instance, if an MTMM analysis demonstrated that two different measurements of the same construct did not correlate well, does that mean they are inaccurate or does that mean that the construct under investigation is multi-faceted in the same way that Bach's fugues are multi-faceted?

The third path – and the one that we recommend – is considering both of these routes as these multiple measurements are considered. In other words, when do our measurements measure the same aspect of a construct and when do they measure different aspects of a construct. For example, although *strategy use* is considered a construct within a domain, different aspects of that strategy use (i.e., quantity, quality, and conditional use) have been demonstrated to be related to learning in different ways (Dinsmore, 2017). Thus, how can we operationalize our theoretical conceptions of strategy use in meaningful ways to build and use theory? This is particularly important as we think about the development (e.g., changes) of these processes as they unfold over time. Like our Tocatta and Fugue in D minor example, it is quite possible that we begin with a simple melody, but then morph into a more complex interweaving of voices as the development of the piece progresses.

5. Empirical contributions

To explore how we can improve self-report measurements or use them in concert with other measurements, eight empirical studies were conducted. These studies examined the validity of score interpretations and future of self-report measurements. These studies each addressed at least two of the Special Issue's three central questions:

1. In what ways do self-report instruments reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use?
2. How does the use of self-report constrain the analytical choices made with that self-report data?
3. How do the interpretations of self-report data influence interpretations of study findings?

Durik and Jenkins's (2020) test of the role of certainty with self-report surveys. They build on literature connecting attitude to behaviour, seeking a new perspective on the relationship between interest and behaviour. This paper tests the relationship between level of interest and certainty of that self-report. This is then extended to examination of the connections between certainty and related behaviour. Durik and Jenkin's is a rare attempt at Vertical innovation with interesting preliminary implications for survey methods and interest research theory. This research needs to be followed up with different participants and variations on their research design.

Chauliac et al. (2020) employed eye-tracking to assess the cognitive processes participants undertake while completing a quantitative questionnaire. They aimed to establish linkages between participants eye movements and their questionnaire answering behaviour. This research yielded no simple answers but lays a foundation for further research into the processes underlying questionnaire response behaviours – namely, in helping to figure out if the questionnaires and eye movements were measuring similar or different aspects of those underlying reading processes.

Making a case for the multimethod approaches that recognise both the value and weakness of self-report, Vriesema and McCaslin (2020), bring survey self-report and observations together in their article. Their results suggest that there is clear alignment between self-report and classroom observations of student groups at ages as young as grade three. Their findings support the use of self-report as part of robust research design for a broad range of ages.

Rogiers et al. (2020) employed think aloud protocols to further explore person centered survey self-report findings regarding secondary school students' text-learning strategies. Results from this combination of retrospective and concurrent approach to self-report pointed to the validity of self-reports. The latter approach provided an additional, nuanced, often ignored perspective on the frequency and sequence of students'

strategies. This article reviews how this pair of self-report methods offers researchers a unique bifocal perspective on student learning experience.

Iaconelli and Wolters (2020) address an area of survey research which is often noticed but rarely engaged with: Insufficient Effort Responding to surveys. This research tests whether "insufficient effort responding" (IER) to survey question is a meaningful threat to survey data validity. As important as their findings, which point to IER as more nuisance than threat, are their recommendations for survey research when reporting their findings.

Toward Vertical innovation of survey self-report in this mobile age, Fryer and Nakao (2020) present an experimental test of four survey formats (Likert, Visual Analogue Scale, Slide, and Swipe). A series of analyses on the resulting data set encourage more work with continuous formats like Slide and Swipe. Nearly a century on from the inception of formats like Likert and VAS, the authors suggest it is time for researchers to look up and embrace our touch-based future.

Van Halem et al. (2020) presented a study triangulating survey self-reports of self-regulated learning with online traces of students learning behaviours. They confirm that aptitude-based self-reports cannot accurately capture complex SRL alone. Their findings suggest that self-report measures and online data regarding SRL are complementary in predicting students' study success. Results demonstrate that both perspectives explain a unique proportion of students' academic performance.

Moeller et al. (2020) aimed to make students' course feedback more meaningful to instructors. They did this through research design and analyses that separate the broader learning situation from the individual's reported experience. This separation makes it possible to track the subjective and objective development of learning experiences across a course. In addition to demonstrating how their methods might support individualised learning, Moeller et al.'s study raised the critical role of multiple methods and including objective measures in self-report centered research.

6. The Commentaries

In addition to these empirical contributions, three international experts have weighed in on how this Special Issue's articles make substantive contributions to the extant research literature. Each focus on at least two of the three guiding questions for the special issue. Winne (2020) takes a conceptual approach by focusing on what self-report data are. While there is much discussion in the literature about our conceptualizations of constructs, there is much less discussion about how we conceptualize the measurements themselves. Winne tackles this thorny issue. Winne argues – and we agree – that without a better conceptualization of self-reports, there is little evidence that participants can get better at responding to them. In turn, the better participants are at responding to these types of measurements, the better the interpretations of these data will be. Pekrun (2020) makes the case for the importance of self-report data, and like Winne touches on what they are. Pekrun extends this discussion by focusing primarily on how to improve the validity of the score interpretations of self-report data. Finally, Van Meter (2020) tackles all three questions. At the heart of her commentary is a deep dive into when self-reports are useful and how they can be leveraged to best help us build and refine theory. Her theoretically-driven set of conditions for when and how to use self-report offer both younger and more experienced researchers alike a useful framework to guide their choices of self-report measurements.

7. Implications of the Special Issue

This brings us full circle back to how the empirical and commentary articles have together addressed this Special Issue's focal points. The eight empirical contributions stretched across the theoretical domains of

self-regulation, interest and cognitive processing strategies, but still presented a coherent picture of the validity and future of self-report.

1. In what ways do self-report instruments reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use?

A common theme across Rogiers et al (2020), van Halem et al. (2020) and Vriesema et al. (2020) is that the retrospective survey self-report of attitudes and dispositions are an important often unique part of understanding future learning experience and outcomes. However, for more comprehensive, dynamic conceptualisations to be drawn, additional online measurement is critical. This online measurement might be self-report (TAP) or observed (trace or observations), both perspectives have the potential to expand our understanding of students' strategies and motivations for learning. The answer to the SI's question is therefore that instruments do matter, and the path toward more robust conceptualizations is multimethod research designs. Any questions about whether those methods should be self-report or not can be set aside.

2. How does the use of self-report constrain the analytical choices made with that self-report data?

The wide variety of contributions to this special issue demonstrate it is the broader question of research design that determines analytical choices as much or more than how self-report is used. Experimental, repeated measures, variable/person-centered analyses and an array of mixed methods arrangements exemplify the full range of analytical tools available to researchers. Researchers are strongly encouraged to focus less on well-known issues with self-report, and instead look to the designs they are embedded within and analyses employed.

3. How do the interpretations of self-report data influence interpretations of study findings?

While each of this Special Issue's articles addresses this question in some form, the syntheses of the three commentaries address it best. All three of these commentaries point out the need to clearly understand the core construct (e.g., Pekrun, 2020), the measurement itself (Winne, 2020), and the conditional nature of their use (Van Meter, 2020). It is critical that the interpretations of self-report data are situated within theoretical frameworks of the core constructs that are being measured. For instance, interpretations of self-report data around interest (e.g., Durik & Jenkins, 2020) will be qualitatively different than those around feedback (Moeller et al., 2020). In other words, the self-report itself must change to allow better interpretations; even survey formats must remain flexible to innovation (e.g., Fryer & Nakao, 2020).

8. Concluding thoughts

The impetus of this special issue was borne out of our frustration as younger scholars with the tools used to study the covert, complex processes at the heart of this special issue. Having seen self-report used poorly and hearing the calls from scholars at all stages of their career calling for a moratorium on self-report, we wanted to expand the discussion beyond simply deciding whether we should forge ahead with the same old tools in the same old way or abandon them all together. Rather, we wanted a vehicle in which scholars could reflect on the way these tools are used and use them more appropriately. We were fortunate to have a number of scholars willing to contribute high-quality empirical studies to this effort. The three commentaries then provided excellent avenues for extending these conversations and hopefully spurring more deep conversations about these thorny issues. We hope that readers of this special issue will be as satisfied with the result as we were in helping to curate them.

9. References

Bondarev, A., & Greiner, A. (2019). Endogenous growth and structural change through vertical and horizontal innovations. *Macroeconomic Dynamics*, *23*, 52-79. https://doi.org/10.1017/S1365100516001115

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Chauliac, M; Catrysse, L. ; Gijbels, D. & Donche V. (2020). It is all in the surv-eye: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*. 8 (3), 26 – 39. https://doi.org/10.14786/flr.v8i3.48

Chiu, M. H., Liaw, H. L., Yu, Y. R., & Chou, C. C. (2019). Facial micro-expression states as an indicator for conceptual change in students' understanding of air pressure and boiling points. *British Journal of Educational Technology*, *50*, 469-480. https://doi.org/10.1111/bjet.12597

Dingle, G. A., Hodges, J., & Kunde, A. (2016). Tuned In emotion regulation program using music listening: Effectiveness for adolescents in educational settings. *Frontiers in Psychology*, *7*, 859. https://doi.org/10.3389/fpsyg.2016.00859

Dinsmore, D. L. (2017). Towards a dynamic, multidimensional model of strategic processing. *Educational Psychology Review, 29,* 235-268. https://doi.org/10.1007/s10648-017-9407-5

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20,* 391-409. https://doi.org/10.1007/s10648-008-9083-6

Durik, A. M. & Jenkins J. S. (2020). Variability in Certainty of Self-Reported Interest: Implications for Theory and Research. *Frontline Learning Research*. 8 (3) 85-103. https://doi.org/10.14786/flr.v8i3.491

Fyer, L. & Nakao K. (2020). The Future of Survey Self-report: An experiment contrasting Likert, VAS, Slide, and Swipe touch interfaces. *Frontline Learning Research*, 8 (3),10-25. https://doi.org/10.14786/flr.v8i3.501

Gillet, N., Morin, A. J., Huyghebaert, T., Burger, L., Maillot, A., Poulin, A., & Tricard, E. (2019). University students' need satisfaction trajectories: A growth mixture analysis. *Learning and Instruction*, *60*, 275-285. https://doi.org/10.1016/j.learninstruc.2017.11.003

Ginns, P., Martin, A. J., & Papworth, B. (2018). Student learning in Australian high schools: Contrasting personological and contextual variables in a longitudinal structural model. *Learning and Individual Differences*, *64*, 83-93. https://doi.org/10.1016/j.lindif.2018.03.007

Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, *65*, 896-928. https://doi.org/10.1111/lang.12136

Hidi, S. (2016). Revisiting the role of rewards in motivation and learning: Implications of neuroscientific research. *Educational Psychology Review*, *28*(1), 61-93. https://doi.org/10.1007/s10648-015-9307-5

Iaconelli, R. & Wolters C.A. (2020). Insufficient Effort Responding in Surveys Assessing Self-Regulated Learning: Nuisance or Fatal Flaw? *Frontline Learning Research*. 8 (3) 104 – 125. https://doi.org/10.14786/flr.v8i3.521

Lawless, K. A., & Riel, J. (2020). Exploring the utilization of the big data revolution as a methodology for exploring learning strategy in educational environments. In D.L. Dinsmore, L. K. Fryer, & M. M. Parkinson (Eds.), *Handbook of strategies and strategic processing*, (pp.296-316). New York: Routledge.

Lawrence, J. G. (2005). Horizontal and vertical gene transfer: The life history of pathogens. *Contributions to Microbiology*, *12*, 255-271.

Kline, R. B. (2011). *Principles and practices of structural equation modeling* (3 ed.). New York: Guilford Press.

Martin, A.J. (2011). Prescriptive Statements and Educational Practice: What Can Structural Equation Modeling (SEM) Offer? *Educational Psychology Review*. *23*. 235-244. https://doi.org/10.1007/s10648-011-9160-0

Mayer, R. E. (2017). How can brain research inform academic learning and instruction? *Educational Psychology Review*, *29*(4), 835-846. https://doi.org/10.1007/s10648-016-9391-1

Moeller, J. ;Viljaranta, J.; Kracke, B. & Dietrich, J. (2020). Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Frontline Learning Research*, *8*(3), 63-84. https://doi.org/10.14786/flr.v8i3.529

Pekrun, R. (2020). Self-report is indispensable to assess students' learning. *Frontline Learning Research*, *8* (3), 185–193. https://doi.org/10.14786/flr.v8i3.627

Rogiers, A.; Merchie, E. & Van Keer H. (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research*, *8*(3), 40 – 62. https://doi.org/10.14786/flr.v8i3.527

Van Halem, N., van Klaveren, C., Drachsler H., Schmitz, M., & Cornelisz, I. (2020). Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data. *Frontline Learning Research*, 8(3) 140-163; https://doi.org/10.14786/flr.v8i3.497

Van Meter, P. (2020) Measurement and the Study of Motivation and Strategy Use: Determining If and When Self-report Measures are Appropriate. *Frontline Learning Research*, *8*(3), 174–184. https://doi.org/10.14786/flr.v8i3.631.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*, 3-14. https://doi.org/10.1007/s11409-006-6893-0

Vriesema, C.C., & McCaslin, M. (2020) Experience and Meaning in Small-Group Contexts: Fusing Observational and Self-Report Data to Capture Self and Other Dynamics. *Frontline Learning Research*, *8*(3), 126-139. https://doi.org/10.14786/flr.v8i3.493

Winne, P. (2020) A Proposed Remedy for Grievances about Self-Report Methodologies. *Frontline Learning Research*. 8 (3) *164 -173.* https://doi.org/10.14786/flr.v8i3.625

Yuen, A. H., Cheng, M., & Chan, F. H. (2019). Student satisfaction with learning management systems: A growth model of belief and use. *British Journal of Educational Technology*, *50*(5), 2520-2535. https://doi.org/10.1111/bjet.12830

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*, 82-91. https://doi.org/10.1006/ceps.1999.1016

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 10 - 25
ISSN 2295-3159

The Future of Survey Self-report: An experiment contrasting Likert, VAS, Slide, and Swipe touch interfaces

Luke K. Fryer[a] & Kaori Nakao[b]

[a]Faculty of Education, The University of Hong Kong, Hong Kong

[b]Seinan Gakuin University, Fukuoka, Japan

Article received 1 June 2019 / Article revised 4 December / Accepted 6 December / Available online 30 March

Abstract

Self-report is a fundamental research tool for the social sciences. Despite quantitative surveys being the workhorses of the self-report stable, few researchers question their format—often blindly using some form of Labelled Categorical Scale (Likert-type). This study presents a brief review of the current literature examining the efficacy of survey formats, addressing longstanding paper-based concerns and more recent issues raised by computer- and mobile-based surveys. An experiment comparing four survey formats on touch-based devices was conducted. Differences in means, predictive validity, time to complete and centrality were compared. A range of preliminary findings emphasise the similarities and striking differences between these self-report formats. Key conclusions include: A) that the two continuous interfaces (Slide & Swipe) yielded the most robust data for predictive modelling; B) that future research with touch self-report interfaces can set aside the VAS format; C) that researchers seeking to improve on Likert-type formats need to focus on user interfaces that are quick/simple to use. Implications and future directions for research in this area are discussed.

Keywords: Likert; VAS; Slide; Self-report; Response format; Experimental design; Mobile; Touch interface

Corresponding author: Luke K. Fryer, CPD 1.80, The University of Hong Kong, Pokfulam Road, Hong Kong.
Email: fryer@hku.hk *DOI:* https://doi.org/*10.14786/flr.v8i2.502*

1. Introduction

The present special issue (Fryer & Dinsmore, 2020) has planted its flag in an unpopular or extremely popular—depending on your perspective—area of research. Unpopular because of the nature of self-reported data: if it is qualitative, it lacks external validity, and if it is quantitative, it is ordinal at best. Unpopular because it is, after all, just intra-psychic "stuff", only loosely tied to the observed, interval/ratio construct gold standards (i.e., our own version of physics envy; see Howell, et al., 2014). Popular because very few researchers in the social sciences can avoid self-report in some form or another. For these reasons, and the fact that year on year humankind collects and analyses more self-report data than ever before, self-report data, and how it is collected, deserves more of our attention.

Although there are many, many means of collecting these self-reports, surveys/questionnaires are the most ubiquitous. Despite the fact that the two most popular formats for surveys have been around for nearly a century (i.e., Visual Analogue Scale VAS[1], Hayes. & Patterson, 1921; Likert[2], Likert, 1932), very little has been done to improve on them. The scant existing research comparing them has often concluded with a statement equivalent to "same difference".

Only during the past two decades has the ground begun to shift under Likert and VAS formats. Computers made slider formats possible, VAS easier to implement and, survey data in all formats far easier to obtain. Mobile devices have led to a natural expansion in the amount of surveys, but little actual development in how they are conducted.

Towards this development the current study presents longstanding issues (commonly and rarely addressed) alongside newer factors made prominent by computer and mobile survey interfaces. This short review is complimented by an experimental study comparing Labelled Categorical Scale (LCS; a Likert format with no center/neutral point), VAS, Slider (a sliding bar with labels and numbers) and a new format/interface (Swipe; an adaptation/extension of the slide format) for collecting quantitative self-reported, micro-analytic data regarding students' interest in classroom tasks.

2. Background

2.1 The criticisms and critical roles of self-report

There are many means of collecting self-reported information, but the most common means of self-report across all fields of human sciences are surveys measuring agreement to a set of statements across a numerical scale of some type (See Durik & Jenkins, 2020). Being the most common type of self-report, surveys also receive the most criticism. These censures generally focus on two critical weaknesses inherent in survey data. The first is the often ordinal (or at least not technically interval) nature of the data itself. The second concern has two related parts, the first is that it is latent and therefore invisible to the senses, the related second part is the data's often tenuous (and generally indirect) connection to the observed world. Nearly every researcher working with survey data has received a review of their manuscript pointing to one or both of these concerns as a limitation – if not as a reason for rejection.

Despite these acknowledged weaknesses, self-reported data are often the only or most direct means (at a large scale) of getting at human psychology. The obvious areas it is critical in assessing are intra-psychic aspects like beliefs, motivations, and emotions. Less obvious, but an equally important area where self-reports are essential tools, are processes which are partially evident to the observer, but like an iceberg are mostly submerged: i.e., metacognitive and cognitive strategies.

In addition to the broad concerns regarding the fact that these data are "just self-reported", there are a host of other issues less often discussed, and often unresolved, with quantitative survey data. The current study presents a brief review of some of these issues ranging from those that are (a) longstanding

and commonly addressed, to (b) longstanding less often discussed, and finally (c) modern issues specific to computer and mobile (touch) interfaces. Following this brief "highlight" review supported by recent research from a range of domains, a short experimental study examining four touch interfaces, with four self-report formats, for collecting survey data through mobile phones immediately after classroom experiences will be presented. Discussion will seek to tie the review and experiment together, while lighting the way for more understanding, research and general development in this critical, but often unquestioned area of research methods.

2.2 Longstanding issues with survey research

Before engaging with less often addressed issues with survey data, two important concerns commonly addressed through design and analyses should be noted. The first is the "less than interval" nature of survey data (i.e., it might be continuous but who knows what "it" is). This problem is generally addressed along with construct validity and reliability by the use of multiple items and either mean- or, preferably, latent-variable analysis. Latent-variable analysis is preferred for a range of reasons, of which measurement error is most commonly cited. Algorithms such as those natively used by latent software packages like M*plus* (Muthén & Muthén, 1998-2015) are purported to ameliorate the stepwise nature of ordinal data, smoothing the distribution that classical statistics relies upon. Reliability is supported by scales utilising items with similar content and reliability that can be assessed at a latent level (Raykov, 2009), offering flexibility to latent modelling research.

2.3 Longstanding less often discussed issues with survey research

Some of the longstanding, but often left unspoken, issues with survey data include central tendency, ceiling effects, number of appropriate categories, influence of proximal items, and self-report agreement vs. magnitude. Central tendencies generally occur when survey respondents over subscribe to middling amounts of agreement and can be related to the use of a non-committal category (Foddy, 1994). While central tendency has long been seen as bias, recent Bayesian analysis suggests it might actually be a reflection of the probability of surveyed choice (Douven, 2018). This is still an unresolved issue and many researchers will no doubt continue to blame scale midpoints as the source of this problem.

Likert format surveys (Voutilainen, et al., 2016) and the survey statements themselves (Austin & Brunner, 2003) have been linked to ceiling effects. Ceiling effects are when a large proportion of survey respondents report the highest possible scale value. Like central tendency biases, ceiling effects can affect the normality of data and result in Type I errors (Austin & Brunner, 2003).

The number of appropriate categories in survey report formats is one of those issues that all researchers have to face when designing instruments and often results in a best guess. Linked with concerns regarding central tendency, these questions also focus on odd vs. even numbers of categories (e.g., Adelson & McCoach, 2010).

The last issue is that of the difference between agreement with survey labels and the magnitude of that agreement (Berger & Alwitt, 1996). A two-step approach, with a Likert response format followed by a cumulative scale from not very strong to very strong has been suggested as a mechanism for assessing both aspects of respondents' experience (Albaum, 1997). While this pairing of self-report has presented robust predictive strength for related variables, this line of research has not been consistently pursued (see Durik & Jenkins, 2020).

2.4 New Issues with survey research

Four relatively new issues that computer and now mobile interfaces are making central are (a) The use and number of labels and/or ticks on a slider or VAS report scale line (no longer focused on explicitly stated categories), (b) Precision in selecting the level of self-report, (c) Speed in selection, (d)

Bias due to everything from age to education, and (e) Relative non-response to different scale formats. For most of these issues there is only a budding body of research to draw upon.

Matejka, et al. (2016) is to our knowledge the only in-depth study testing the effect of the number of ticks (on a slide line) on self-report precision and speed. This study indicated that with regard to precision that 11 ticks is superior to five. This study also supported the use of dynamic feedback (a running quantitative score above the moveable slide marker). This addition enhances precision but has a detrimental effect on the speed of self-report. This study also pointed to the benefit of banded coloring of the slider line to signify increments as being superior to ticks alone.

Bias is a complex area to research and has not to our knowledge been properly investigated with studies supported by experimental design. Survey studies have noted apparent biases supporting Labelled Categorical Scale (LCS; i.e., Likert-type) interfaces over Slide and VAS interfaces (Voutilainen et al., 2016). This research attributes the benefits of LCS to age (i.e., easier for older and younger respondents) and/or education (i.e., easier for respondents with less education). Their very specific supposition regarding bias reflects broad support for LCS over other continuous survey interface formats.

2.5 Four formats for self-reporting agreement: LCS, VAS, Slide, and Swipe

A considerable number of studies, in a wide range of domains have assessed the relative usefulness of different survey interfaces. The majority have focused specifically on LCS and VAS, which have been the predominant self-report formats. On comparing LCS and VAS, most studies conclude that they are highly correlated and present similar overall distributions of data (Bolognese et al., 1990; Reed, et al., 2017; Vickers, 1999). If we include ease of administration, these studies generally support the use of LCS over other formats (i.e., generally VAS).

A smaller number of studies comparing LCS and VAS have cited similar consistencies between the two-survey format but fallen on the VAS side of the fence. These studies often cite the interval nature of the VAS data relative to the ordinal nature of LCS data (Bishop & Herron, 2015). Some of these studies also note VAS' robustness to ceiling effects and, in some cases, shorter time to complete when compared to LCS (Couper, et al., 2006; Voutilainen et al., 2016). Lower standard deviation for VAS vs. LCS has been reported, but has been difficult to replicate (Kuhlmann, et al., 2017).

As more surveys go online there has been a related increase in research examining slider interfaces as self-report tools. This research is generally focused on specific aspects of sliders, rather than comparing them to VAS or LCS (radio button) interfaces. What little comparative research there is has suggested no significant differences between Slider and LCS response formats (e.g., Roster, et al., 2015). Research has also pointed towards non-response being higher for Slider compared to LCS response formats (Liu, 2017). Research of specific Slider related issues such as direction (Liu, 2017), suggest that the direction of the labels has no effect on self-report outcomes. The starting values for Slider markers have an impact on 101, but not 21 or seven-point scales. However, forcing users to click the scale to start (i.e., no marker initially visible), increases missing data, particularly for 101-point scales (Liu & Conrad, 2018).

The present review provides scant clear direction for continued research in the area of survey responses. The majority of the studies presented have pursued a relatively weak research design (Liu and colleagues' programme is a nice exception to this problem) and focused exclusively on longstanding 20th century approaches to survey response formats (LCS and VAS). In line with more recent research focused on computer-based surveys and the growing use of sliders, the future of self-report (like almost everything else) is mobile and touch based. It is critical that as our medium for engaging with media changes, that we adapt the ways in which we structure these media. For example, researchers should consider why we often use a touch radio button when something more intuitive and potentially more powerful might be invented. As Wetzel and Greiff (2018) have called for, future research needs to seek alternative response formats. In the current study we therefore pursued an experimental approach (i.e., random assignment of a three-item survey's scale interface) to testing both well-known and new response formats on mobile touch-based devices.

2.6 An empirical test of four mobile interfaces for survey data collection

Four survey interfaces were compared: Labelled Categorical Scale (Likert-type), VAS, Slider, and Swipe. The first two were included due to their prevalent use across the previous century of research. The third (Slider) was included because of its increasing use through computer and now mobile devices. The fourth (Swipe; Fryer & Fryer, 2019) was included to test some new and alternative approaches that touch interfaces afford. Swipe is built on a basic Slider interface, but is presented on a slope. Users "swipe" along the 45-degree angle (up, left to right) to move a ball up the incline. Consistent with Matejka et al. (2016), this interface integrated dynamic feedback and an approach to banding the intervals between labels in addition to ticks. Ticks were presented both for the six labels and at 1/10 increments between the labels.

3. Aims, Research Questions and Hypotheses

In the current study we aimed to highlight established issues with survey data, some of which are regularly addressed, and others less often discussed. In the current study we also aimed to introduce new questions that survey measurement faces as it integrates with the digital, increasingly mobile age. Embracing this mobile era of survey use, the current study concludes with a brief experimental study comparing the four survey interfaces: LCS, VAS, Slider and Swipe.

Five research questions (RQ) were addressed in the current study's experiment. Sufficient prior research existed to support a hypothesis for one of the questions, the lower time to complete for LCS (Likert type in most cases) relative to other interface formats. First, we were interested in whether the reliability (Cronbach's Alpha) for scales would vary meaningfully across the four self-report interfaces (RQ1). Second, we aimed to determine whether any mean differences in interest for each of the six tasks separately could be attributed to the four interfaces (RQ2). Third, we aimed to assess/compare the predictive relationships from (a) prior interest and self-efficacy to the task interest (with each interface) and then (b) from the task interest to future interest in the course and domain (RQ3). Fourth, we sought to assess and compare the latent structure of the interest constructs measured by each of the four interfaces (RQ4). Fifth, potential differences in central tendency of the data resulting from each interface were compared, looking for patterns of response bias that might be due to the four interfaces (RQ5). Finally, we were interested in whether the time to complete the surveys varied meaningfully across the four interfaces. In this case, we hypothesised that Likert would be the fastest to complete (Hypothesis-1).

4. Methods for the interface comparison

4.1 Participants, Ethics and Procedures

Participants for the current study were postgraduate students (n = 81; Female = 38; resulting in 644 responses) from one research intensive university in Hong Kong. Students came from eight of the university's 10 faculties. Students were completing a short course in preparation for teaching responsibilities as a part of their degree.

The comparison of the four interfaces (survey formats) was undertaken within a broader project examining students' interest in course tasks, the course itself, and the domain of teaching and learning. Across the course, participating students responded to short surveys either directly after tasks (task interest) or at the beginning/end of the course (course and domain interest). All surveys were completed during regular class time.

Students completed the short surveys on their mobile phones by capturing a QR code (embedded in course power points) which directed them to a survey within a custom designed online platform for micro-analytic surveys. The survey interface students engaged with were randomised for each survey QR code, meaning that students had an equal chance of facing any of the four interfaces for each of the six task interest surveys they were asked to complete. In the current study we therefore pursued a within-

individual experimental design. As the interfaces were randomised for each of the six surveys, there was no guarantee that students would engage with all four interfaces and even if, by chance, they did, the number could not be even. This means that this study also relied on between student differences as well.

For the main component of the current study (a comparison of four quantitative self-report interfaces) a three-item survey designed to assess students' interest in a specific task was utilised: This activity/task is personally meaningful; This activity/task is interesting; I want to learn by doing more activities/tasks like this. The predictive validity for the task scales were tested using regression from a future course interest scale consisting of five items (i.e., This course is personally meaningful; This course is interesting; I want more courses like this; I'm enjoying learning about teaching during this course; This course stimulated my curiosity about teaching) and a domain interest scale consisting of five items (i.e., 1) How much do you know about teaching?; In your spare time, how often have you tried to learn about teaching?; I have spent time learning about teaching on my own. How well does this statement match you?; I'm confident in my knowledge of teaching. How well does this statement match you?; I always have questions about teaching. How well does this statement match you?). All survey items were self-reported across a scale 0-5. Labels for Task, Course and Domain (3, 4, 5) asked students to what degree the item matched them specifically (Not at all 0 - Completely 5). Domain item 1 used the labels Almost Nothing 0 - Almost Everything 5. Domain item 2 used the labels Almost Never 0 - Almost Always 5. The task and course scales have demonstrated acceptable reliability and construct validity in several past uses (Fryer, et al., 2020; Fryer, et al., 2019; Fryer, et al., 2017; Fryer, et al., 2016). The domain-level, depth of interest scale was developed recently (Renninger, & Schofield, 2014). It is consistent with current conceptions of individual interest and its development (i.e., Renninger & Hidi, 2015).

In preparation for the current study, ethical approval was sought and obtained from the University's Human Research Ethics Committee (Ethics Approval #1608028). Prior to beginning the study, all students read an overview of the project, were informed that their self-reports would be anonymous and invited to contribute their self-reports to the research project. Six students declined to participate in the research after reading the ethics statement and were removed from the current study, resulting in the aforementioned n-size.

4.2 Analyses

Analysis for the empirical component of the current study began with an examination of the overall and interface specific descriptive statistics for the scale means and reliability (RQ1). ANOVA were conducted for the four interfaces, overall and task by task (RQ2). Regressions were conducted from prior Domain interest and Course self-efficacy predicting Task interest for each interface separately; then regressions from Task interest predicting Course interest in the future was conducted and compared (RQ3). Factor loading (Confirmatory Factor Analysis) for results from each interface was then compared (RQ4). The central tendency, both visual and Skew/Kurtosis were calculated and reviewed (RQ5). Finally, an ANOVAs were conducted to compare completion times for each interface (Hypothesis #1).

5. Results

5.1 Descriptive statistics and Reliability

The overall means for the four interfaces (across all six tasks), for the pre-post measure, their differences by task, and scale reliabilities are presented in Table 1. The reliability for each scale and for the task interest scales used with each of the four interfaces were all well above what is commonly suggested as being acceptable (> .70; Devellis, 2012). Significant differences were observed for the four interfaces across the project as a whole, but at each of the individual tasks assessed for interest no statistically significant differences were found (Table 1 & 2).

Table 1
Means, Cronbach's Alpha and ANOVA for the four interfaces across all tasks

	LCS	Slide	Swipe	VAS	Prior Course self-efficacy	Prior Domain Interest	Post Course Interest	Post Domain Interest
Means across all tasks	3.21	3.50	3.18	3.60	1.82	1.66	3.64	3.45
Cronbach's Alpha	0.95	0.92	0.96	0.93	0.89	0.78	0.94	0.93
SD	1.07	0.93	1.11	0.98	0.87	0.70	0.95	0.88
p	<.0001							
F	7.21							
n	644							
DF	3							
R^2	0.03							

Table 2
ANOVA for differences for each task

	Task Interest a	Task Interest b	Task Interest c	Task Interest d	Task Interest e	Task Interest f
LCS	3.5	3.04	3.74	2.9	3.21	3.34
Slide	3.3	3.42	3.43	3.6	3.43	3.66
Swipe	3.85	2.82	3.35	2.96	3.59	3.25
VAS	3.69	3.72	3.73	3.81	3.62	3.93
p =	0.08	0.13	0.64	0.06	0.50	0.24
F	2.36	1.9	0.56	2.55	0.79	1.32
n	81	60	66	57	56	62
DF	3	3	3	3	3	3
R^2	0.08	0.09	0.03	0.12	0.04	0.06

Note: Task Interest means for Tasks a-f for each self-report format

4.2 Predictive difference by interface

Regression was used to experimentally test (i.e., random assignment of interface) the prediction from prior domain interest and perceived self-efficacy for the course to the four survey interfaces used for all the six tasks (combined). This test was then followed by regression predicting future interest in the course and domain from students' interest in the course tasks, again for each of the task interest survey interfaces (Table 3). For the prediction from prior domain interest to future tasks, the R2 (.04) was consistent for all but VAS, which presented a non-significant ($p < .05$) relationship. Course self-efficacy was significant for all four interfaces presenting the highest R2 the new interface (swipe = .09) and the lowest for VAS (.06). Task interest predicting future interest in the course (generally strong in past research with these constructs: e.g., Fryer, et al., 2019; Fryer, et al., 2017; Fryer, et al., 2016) resulted in substantially more variance being explained (R2) (Slide = .49, Swipe = .37, VAS = .29, LCS = .28). A similar pattern of relationships resulted for tasks predicting future domain interest (Slide = .51, Swipe = .37, LCS = .35, VAS = .33).

Table 3
Regression Findings

		Interfaces			
Predicted by		LCS	Slide	Swipe	VAS
Prior Domain Interest	*p*	=0.01	<0.001	=0.01	=0.08
	R^2	0.04	0.04	0.04	0.02
	n	179	143	166	147
Prior Self-efficacy	*p*	<0.001	=0.03	<0.001	=0.004
	R^2	0.08	0.062	0.091	0.060
	n	179	143	166	147
Predicted					
Future Domain Interest	*p*	<0.0001	<0.0001	<0.0001	<0.0001
Future Domain Interest	R^2	0.35	0.51	0.37	0.33
	n	158	123	139	124
Future Course Interest	*p*	<0.0001	<0.0001	<0.0001	<0.0001
Future Course Interest	R^2	0.28	0.49	0.37	0.29
	n	137	111	126	116

Note: n refers to the number of survey completions

4.3 Confirmatory Factor Analytic Loading each item for each interface

The CFA loading findings were generally consistent across the four interfaces (Table 4). "Interesting" generally presented the strongest loading, followed by a desire to reengage and finally perceptions of tasks being personally meaningful.

Table 4.
CFA loading for each item for each survey question format

	LCS	Slide	Swipe	VAS
Personally Meaningful	0.51	0.70	0.60	0.51
Interesting	0.98	0.91	0.97	1.05
Want to do the task again	0.91	0.93	0.90	0.79

4.4 Time to complete differences by interface

The average time to complete the task surveys with each of the four interfaces was calculated and compared (Table 5). Despite the relatively large mean differences, no statistically significant differences were observed ($p < .05$). This is likely due to the relatively large standard deviation for the means.

Table 5
Average time to complete task surveys with each of the four interfaces

Interface	Mean	N	SD
Labelled Categorical	26.65	178	37.34
Slide	30.21	156	17.43
Swipe	37.10	171	60.11
VAS	30.00	152	15.49

Note: n refers to the number of survey completions with the specified format

Central Tendency

Table 6 presents the distribution for the four tested interfaces. Skew and kurtosis for each of the interfaces were within even the strictest heuristics (+1 – -1). The graphical distribution for each survey interface is included in the Appendices (Figures 1-4). The distribution presented by these charts makes visually clear the inherent differences between the types of data the different interfaces result in. VAS presents the most skew and LCS appears to encourage students to choose the same ordinal rank, regardless of question, resulting in large amounts of twos, threes and fours but far fewer scores in between. Swipe and Slide presented the most normal looking distributions.

Table 6
Distribution for the four interfaces

	LCS	Slide	Swipe	VAS
SD		0.93	1.11	0.98
Variance	1.13	0.87	1.24	0.96
Skewness	-0.43	-0.53	-0.49	-0.74
Kurtosis	-0.08	0.48	-0.26	0.56

5. Discussion

A brief review of the extant research in the area of quantitative survey self-report formats (or interfaces in the current context) was presented. The literature reviewed came from a broad range of fields, with much of it providing scant direction beyond support for LCS (commonly Likert in format) due to its ease of administration and in some cases for VAS due to the nature of the resulting data (i.e., interval-like). More recent research examining Slider formats has resulted in a handful of incremental suggestions for the field (e.g., use of dynamic response and potential of banding rather than ticks on the slide area) which have not yet been meaningfully taken up by the field. Some of these findings were integrated into the interface tested alongside LCS, VAS and Slide, in a touch-driven format tentatively named Swipe (an early version of Fryer & Fryer, 2019).

In the short experimental study undertaken, six research questions were addressed. Reliability for each of the interfaces was acceptable, with Swipe and LCS presenting the highest reliability across the 6 tasks (RQ1). No statistically significant mean differences were found for the individual tasks (Table 2), but a significant difference across all tasks was observed—albeit with a small R2 (Table 1). In this case VAS presented the highest overall mean and slide the lowest (RQ2). Predictive modelling was undertaken – prior self-efficacy for the course and interest in the domain predicting future course interest; Task interest predicting future Course and Domain interest – for each of the four interfaces. The clearest contrast was for Task to future Course and Domain interest, where Slide and then Swipe presented the strongest relationship (RQ3). Confirmatory Factor Analysis followed, focusing on item loading, with results suggesting a consistent pattern of loading across the four interfaces (RQ4).

Central tendency for the responses were examined statistically and reviewed graphically (Appendices: Figures 1-4). Skew and Kurtosis were within acceptable boundaries for all four interfaces. Graphical representations of the four distributions suggested that the Swipe interface presented the most normal distribution (RQ5). The time to complete the three-question survey with the four interfaces was compared, indicating that, consistent with our hypothesis, the LCS format was the fastest to complete (but not statistically significant, $p < .05$) (Hypothesis #1). A careful review of the data across the six surveys suggest that the differences between the LCS and the other formats declined precipitously with increased use suggesting a learning effect (i.e., getting used to the new interface) across students' engagement with the task interest self-reports.

5.1 Implications for measurement

Assuming the sample size was large enough for the experimental nature of the study (i.e., random distribution of the four conditions), two general findings standout. The first is the relative predictive strength of responses with each of the four interfaces. Slide, followed by the new interface Swipe, stood out as presenting the strongest βs for future interest in the course and domain. Given the fact that much of the research with surveys like this will be aimed at predictive modelling, this finding is both alarming and potentially hopeful: Alarming as the results suggest that the interface matters and can result in substantial differences; hopeful because it suggests that the Slide and Swipe (i.e., interactive and touch-based) formats have significant advantages over older formats like LCS and VAS.

The second is the time difference to complete the survey. Marked differences supporting past findings pointing to the ease of LCS over other formats like VAS were observed. Rather than suggesting, as many previous researchers have, that LCS is therefore preferred due its ease of administration, we suggest that the flexible nature of mobile devices might be channeled to overcome this issue. A careful review of the survey completion time data suggested that the difference between LCS and Slide/Swipe narrowed substantially with increasing use. More intuitive interfaces for Slide/Swipe might be developed to close the gap, and animated directions for interacting with the interfaces might also be used to ameliorate this issue.

While the skew and kurtosis outcomes for each interface were within acceptable boundaries, the graphical presentation made a clear case for Swipe, VAS and Slide (in that order) as providing a more normal distribution of scores. Given the reliance of most of our statistical procedures on such a distribution and the amount of potential data collected with mobile interfaces in the years to come, it seems reasonable to continue to develop continuous self-report interfaces.

6. Limitations and Future Directions

Despite the experimental design, this study faced a number of limitations that should be addressed by future studies in this area. The first is the learning effect that is apparent for all of the continuous interfaces, but most obvious for the newest version (Swipe). The high SDs that resulted, clearly affected the study's power to detect differences between the interfaces which were apparent in the means but were not statistically significant. In this study only participants' responses to a very short survey were examined, whereas most surveys are much longer. Future studies should examine what effect prolonged survey engagement has on different touch interface experiences as well. While more than 600 individual responses across the four interfaces were collected for this study, the actual sample of participants was quite small and very specific. It is important that future studies embrace a broader sample as well as a larger one.

Implicit in the study's design, analyses were conducted between persons, but participants were represented at multiple time points. This design therefore violates the assumption of independent errors as some of the data is nested within-person. To achieve the sample size necessary for a meaningful experimental test of all four interfaces, this limitation could not be avoided. A future experiment in a more controlled context (rather than a classroom setting) could undertake to obtain a clear counter-balanced sample and avoid this limitation.

This was the first published test of the Swipe interface (a pilot version). This test suggested both positive (high βs and reliability) and negative findings (high SDs and time to complete) for the new interface. Future studies from our research programme continue to refine this approach to self-report. The most recent version of the interface (Fryer & Fryer, 2019) is prefaced by an animated user interface infomercial (to spell out how to interact with it). Additional tests comparing Swipe with the Slide (highest βs) and LCS (fastest time to complete) are being conducted towards fine-tuning this new dynamic, touch-based survey interface.

It is critical to note that the present study's questions focus on students' self-reported emotions, beliefs and desires. It is reasonable therefore to constrain the implications of our results to the use of similar types of survey questions.

One means of continuing to advance the research presented here would be through pairing Think-aloud protocols with survey use (e.g., Chauliac, et al., 2020; Rogiers, et. al., 2020). This would provide a small window into the user's mind, suggesting how/whether a specific self-report format and touch interfaces interact with the self-report experience and outcome: i.e., send in a spider to catch the fly.

An additional important area for investigation is that of surveys which enable the seamless integration of both categorical choice and continuous magnitude. To some degree the Swipe interface sought to combine these elements into a single experience. Future interfaces might extend this work or

separate them into an intuitive two-step process: choose a label and then indicate the strength of your feeling for that category (see Durik & Jenkins, 2020).

7. Conclusions

VAS and Likert-type (LCS) formats are approaching their centenary. At the same time humankind sprints towards touch-based mobile devices as a critical nexus for interacting with and managing its world. Self-report is therefore ripe to be improved (disrupted?). The revolution must start with those of us that rely on surveys for research. Better, easier measurement means clearer results and more of them. As one baby step towards this revolution, results from the present study suggest that VAS might be set aside as an option. It presented no clear benefits over the other interfaces in any of the tests and lacks any clear path to enhancement. In contrast, the present research suggests that continuous and interactive formats (Slide & Swipe) are a strong base for development in this area. The field is waiting for researchers with a penchant for disruptive improvement.

Notes:
1. Visual Analogue Scale (VAS) is "a testing technique for measuring subjective or behavioral phenomena (as pain or dietary consumption) in which a subject selects from a gradient of alternatives (as from "no pain" to "worst imaginable pain" or from "every day" to "never") arranged in linear fashion". (Merriam Webster, 2019)
2. A Likert scale is a "rating system used in questionnaires, that is designed to measure people's attitudes, opinions, or perceptions. Subjects choose from a range of possible responses to a specific question or statement; responses typically include "strongly agree," "agree," "neutral," "disagree," and "strongly disagree." Often, the categories of response are coded numerically, in which case the numerical values must be defined for that specific study, such as 1 = strongly agree, 2 = agree, and so on." (Britannica, 2019)

Keypoints

- The two continuous interactive interfaces (Slide & Swipe) yielded the most robust data for predictive modelling.
- Future research with touch self-report interfaces can ignore VAS formats.
- Researchers seeking to improve on Likert-type formats need to focus on UI that are quick and reliable to use.
- Review of the existing research generally suggests that Likert-type is superior to VAS due to its ease of use.
- Many researchers still maintain that VAS formats yield more robust data than Likert-type formats

Acknowledgements

We would like to acknowledge the contribution of Alex Shum for carefully reviewing a previous draft and his overall contribution to this ongoing project. We would also like to acknowledge Ada Lee and Peter Lau who were central to collecting the data for this research and the broader programme.

8. References

Adelson, J. L., & McCoach, D. B. (2010). Measuring the Mathematical Attitudes of Elementary Students: The Effects of a 4-Point or 5-Point Likert-Type Scale. *Educational and Psychological Measurement, 70*(5), 796-807. https://doi.org/10.1177/0013164410366694

Albaum, G. (1997). The Likert scale revisited. *Market Research Society. 39*(2), 1-21. https://doi.org/10.1177/147078539703900202

Austin, P. C., & Brunner, L. J. (2003). Type I error inflation in the presence of a ceiling effect. *The American Statistician, 57*(2), 97-104. https://doi.org/10.1198/0003130031450

Berger, I., & Alwitt, L. F. (1996). Attitude conviction: a measure of strength and function. *Unpublished paper*.

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International journal of exercise science, 8*(3), 297-302.

Boognese, J. A., Schnitzer, T. J., & Ehrich, E. (2003). Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. *Osteoarthritis and Cartilage, 11*(7), 499-507. https://doi.org/10.1016/S1063-4584(03)00082-7

Britanica.com (2019). Likert definition. Retrieved on November 18, 2019 from https://www.britannica.com/topic/Likert-Scale

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review, 24*(2), 227-245. https://doi.org/10.1177/0894439305281503

Chauliac, M., Catrysse, L., Gijbels, D., & Donche V. (2020). It is all in the surv-eye: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*. 8(3), 26 – 39. https://doi.org/10.14786/flr.v8i3.489

Devellis, R. F. (2012). *Scale development: Theory and application*. New York: Sage

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review, 25*, 1-9. https://doi.org/10.3758/s13423-017-1344-2

Durik, A. M., & Jenkins, J. S. (2020). Variability in certainty of self-reported interest: Implications for theory and research. *Frontline Learning Research, 8*(2) 86-104. https://doi.org/10.14786/flr.v8i3.491

Foddy, W. (1994). Constructing questions for interviews and questionnaires: Theory and practice in social research. Cambridge: Cambridge university press.

Fryer, L. K., Thompson, A., Nakao, K**.,** Howarth, M., & Gallacher, A. (2020). Supporting self-efficacy beliefs and interest as educational inputs and outcomes: Framing AI and Human partnered task experience. *Learning and Individual Differences*. https://doi.org/10.1016/j.lindif.2020.101850

Fryer, L. K., & Dinsmore D.L. (2020). The Promise and Pitfalls of Self-report: Development, research design and analysis issues, and multiple methods. *Frontline Learning Research*, *8*(3), 1–9. https://doi.org/10.14786/flr.v8i3.623

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior, 93*, 279-289. https://doi.org/10.1016/j.chb.2018.12.023

Fryer, L. K., & Fryer, K. (2019).情報処理装置、情報プログラムおよびこれを記録した記録媒体、ならびに情報処理方法.. Patent # 6585129 (Japan).

TRANSLATION: [Dynamic touch based interface for survey self-report; Translation of Japanese patent title: information processor (information technology equipment), information program and a medium for the recording, and a method of information processing]

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior, 75*, 461-468. https://doi.org/10.1016/j.chb.2017.05.045

Fryer, L. K., Ainley, M., & Thompson, A. (2016). Modelling the links between students' interest in a domain, the tasks they experience and their interest in a course: Isn't interest what university is all about? *Learning and Individual Differences, 50*, 157-165. https://doi.org/10.1016/j.lindif.2016.08.011

Hayes, M. H., & Patterson, D. (1921). Experimental development of the graphic rating method. Psychological Bulletin, 18, 98-107.

Howell, J. L., Collisson, B., & King, K. M. (2014). Physics envy: Psychologists' perceptions of psychology and agreement about core concepts. Teaching of Psychology, 41, 330-334. https://doi.org/10.1177/0098628314549705

Jaeschke, R., Singer, J., & Guyatt, G. H. (1990). A comparison of seven-point and visual analogue scales: data from an andomized trial. *Controlled clinical trials, 11*, 43-51. https://doi.org/10.1016/0197-2456(90)90031-V

Kuhlmann, T., Dantlgraber, M., & Reips, U.-D. (2017). Investigating measurement equivalence of visual analogue scales and Likert-type scales in Internet-based personality questionnaires. *Behavior Research Methods, 49*, 2173-2181. https://doi.org/10.3758/s13428-016-0850-x

Likert, R. (1932). "A Technique for the Measurement of Attitudes". Archives of Psychology, 140, 5-55.

Liu, M. (2017). Labelling and direction of slider questions: Results from web survey experiments. *International Journal of Market Research, 59*, 601-624. https://doi.org/10.2501/IJMR-2017-033

Liu, M., & Conrad, F. G. (2018). Where Should I Start? On Default Values for Slider Questions in Web Surveys. *Social Science Computer Review*, *37*(2), 248-269. https://doi.org/10.1177/0894439318755336

Chauliac, M., Catrysse, L., Gijbels, D. and Donce, V. (2020). It is all in the *surv-eye*: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*. 8 (2), 26 – 39. http://doi.org/10.14786/flr.v8i3.489

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). *The effect of visual appearance on the performance of continuous sliders and visual analogue scales.* Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Merriam-Webster. (2019). Visual Analogue Scale definition. Retrieved on November 18, 2019 from https://www.merriam-webster.com/dictionary/likert

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide.* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development, 42*, 223-232. http://doi.org/10.1177/0748175609344096

Rogiers, A.; Merchie, E. & Van Keer (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research*. 8(3) 40 – 62. http://doi.org/10.14786/flr.v8i3.527

Reed, C. C., Wolf, W. A., Cotton, C. C., & Dellon, E. S. (2017). A visual analogue scale and a Likert scale are simple and responsive tools for assessing dysphagia in eosinophilic oesophagitis. *Alimentary Pharmacology & Therapeutics, 45*, 1443-1448. https://doi.org/10.1111/apt.14061

Renninger, K., & Hidi, S. (2015). The power of interest for motivation and engagement. New York: Routledge.

Renninger, K., & Schofield, L. S. (2014). Assessing STEM interest as a developmental motivational variable. Paper presented at the American Educational Research Association, Philadelphia, PA.

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice, 11*(1), 1.

Vickers, A. J. (1999). Comparison of an ordinal and a continuous outcome measure of muscle soreness. *International Journal of Technology Assessment in Health Care, 15*, 709-716. https://doi.org/10.1017/S0266462399154102

Voutilainen, A., Pitkäaho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing, 72*, 946-957. https://doi.org/10.1111/jan.12875

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales: Why we should think more carefully about the response format in questionnaires. *European Journal of Psychological Assessment, 34*, 1-5. http://doi.org/10.1027/1015-5759/a000469

9.Appendices

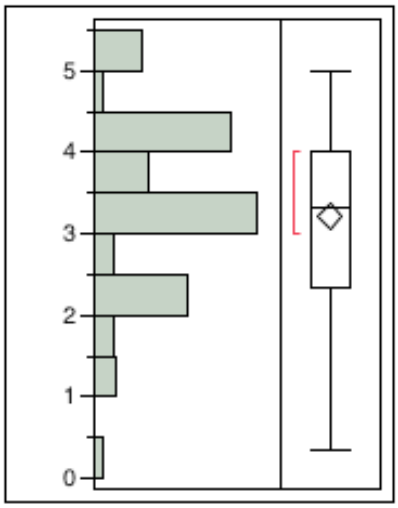


Figure 1. Distributions for Interface for LCS

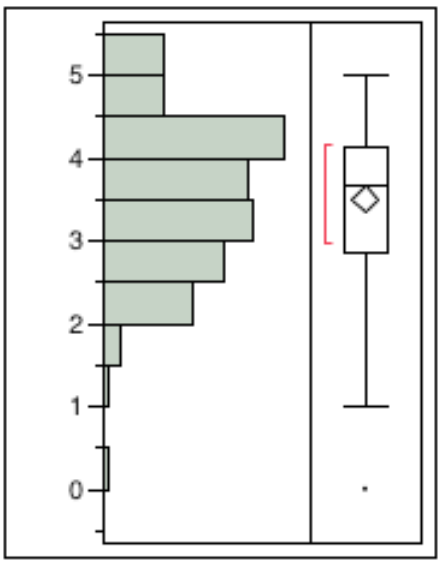


Figure 2. Distributions for Interface for Slide

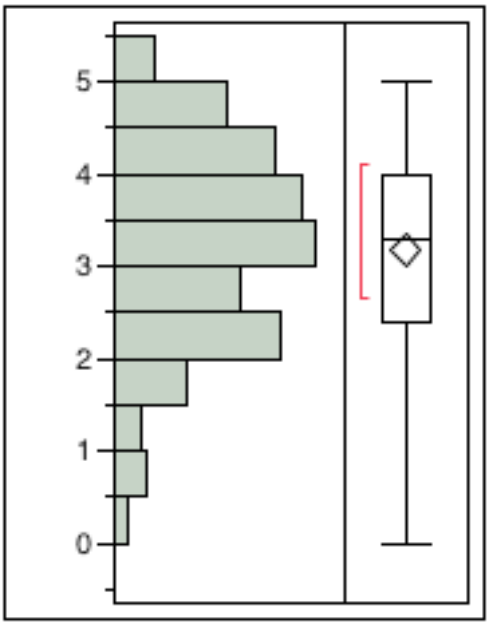


Figure 3. Distributions Interface Swipe

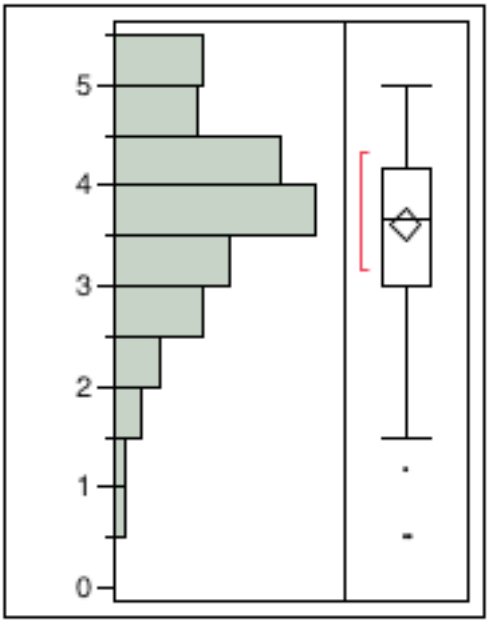


Figure 4. Distribution for VAS Interface

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 26 - 39
ISSN 2295-3159

It is all in the *surv-eye*: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies?

Margot Chauliac[a], Leen Catrysse[a], David Gijbels[a], Vincent Donche[a]

[a]University of Antwerp, Belgium

Article received 13 May 2019 / Article revised 18 February 2020 / Accepted 23 February / Available online 30 March

Abstract

Although self-report questionnaires are widely used, researchers debate whether responses to these types of questionnaires are valid representations of the respondent's actual thoughts and beliefs. In order to provide more insight into the quality of questionnaire data, we aimed to gain an understanding of the processes that impact the completion of self-report questionnaires. To this end, we explored the process of completing a questionnaire by monitoring the eye tracking data of 70 students in higher education. Specifically, we examined the relation between eye movement measurements and the level of internal consistency demonstrated in the responses to the questionnaire. The results indicated that respondents who look longer at an item do not necessarily have more consistent answering behaviour than respondents with shorter processing times. Our findings indicate that eye tracking serves as a promising tool to gain more insight into the process of completing self-report questionnaires.

Keywords: eye tracking; cognitive processes; survey research; self-report questionnaires; working memory capacity

Corresponding author: Margot Chauliac Sint-Jacobstraat 2 S.Z.309 , 2000 Antwerpen
*Margot.Chauliac@uantwerpen.be**, DOI: https://doi.org/10.14786/flr.v8i3.489*

1. Introduction

In self-report questionnaires respondents are asked to answer questions about themselves and as an instrument they are widely used to measure beliefs, attitudes, feelings and opinions in diverse fields of research (Singleton & Straits, 2009). This also holds for the domain of research on learning and instruction where self-report questionnaires are often used to map student learning. Important assets of these questionnaires are that they are easy to administer in both small and large groups and that their use is time and cost-effective. However, despite the reliability, validity and advantages self-report questionnaires might offer, a critical stance towards their use is required to gain more insight into students' cognitive processing strategies (Dinsmore & Alexander, 2012). Many researchers argue that respondents are, consciously or unconsciously, not always able to respond accurately to these questions (Schellings, 2011; Schellings & Van Hout-Wolters, 2011). This inability may influence the consistency by which a respondent scores the different items of a questionnaire, and thus the reliability of its outcomes (Richardson, 2004, 2013; Veenman, 2011; Veenman & van Hout-Wolters, 2005).

In order to assess the quality of the retrieved data, it is critical to examine whether the responses to questionnaires are valid representations of respondents' actual thoughts and beliefs (Schwarz, 2007; Tourangeau et al., 2000). Thus, when a specific set of items focuses on the same topic (i.e. a scale mapping a specific belief), the responses to these items need to be representative of the respondent's beliefs. An individual answering pattern on a set of items that is consistent with the underlying scale leads to reliable survey data. When this is not the case, one can start questioning how the respondent completed the survey and to what extent this is related to the consistency of their answers.

Generally, there is a black box concerning the processes in participants' completion of self-report questionnaires. Gaining an understanding of these processes could help to provide additional insight into the quality of survey data. However, this area of interest, and in particular, the process of completing the questionnaires, has been under-examined in the literature so far. In this study, we use eye tracking to examine the processes at play when completing self-report questionnaires that aim to map students' cognitive processing strategies. In particular, we focus on the specific strategies that students use while processing items.

2. Theoretical Framework

2.1. Cognitive processing when completing self-report questionnaires

Surveys have a long history in educational research (Marsden & Wright, 2010; Rossi et al., 1983). Despite its long history, it is only from 1980 onwards that cognitive psychology started to enter the field of survey research. The focus shifted from the outcomes of the questionnaire to the cognitive processing activities that were at play when completing questionnaires (Fowler, 2014; Willis & Miller, 2011). However, despite the development of the cognitive aspects of survey methodology (CASM), the focus was still on examining how cognitive processes could influence the outcomes of the questionnaires, instead of investigating how the underlying processes while completing self-report questionnaires could be related to the reliability of their outcomes.

Following the CASM-movement, multiple theoretical models were developed to grasp the processes at work in the reading of questions and providing answers to these questions (Jobe & Herrmann, 1996): these included the four-stage model by Tourangeau (1984), the autobiographical question-answering model from Schwarz (1990), the flexible processing model by Willis et al. (1991) and the information processing model of self-report item response (Karabenick et al., 2007). All these models share the common feature of attempting to grasp the complexity of completing survey questionnaires by distinguishing important stages that respondents go through in order to generate an

answer. Most researchers agree that respondents must give meaning to a question and be able to retrieve necessary information from their memory. Only then they will be able to make an informed decision and choose a congruent response option (Karabenick et al., 2007; Tourangeau, 1984). All these models are characterised by the fact that the described stages do not have to follow each other in a linear sequence. One can move back and forth so that there may be iterations and overlap between the steps. It is even possible that one or some of the steps are weakly conducted or completely missing. Nevertheless, one can only expect substantive answers when respondents thoroughly conduct all cognitive processes when answering a question (Krosnick, 1991; Krosnick & Alwin, 1987).

In the field of survey research, the comprehension of the question is an important prerequisite for achieving meaningful results. A crucial step is, therefore, to design a questionnaire such that all respondents understand the items in the same way as the researcher intended (Neuert, 2016). Previous research already demonstrated that comprehension problems could arise or that respondents may satisfice while completing the survey (Krosnick & Alwin, 1987). Another important factor that plays a role is the capacity of a respondent's working memory. Working memory concerns the limited amount of information that can be processed and temporarily stored in the memory while performing complex cognitive tasks (Baddeley & Hitch, 1974). Krosnick (1991) argues that working memory is limited and that respondents are unable to give the latter options as much attention as the ones they consider initially. Moreover, respondents may differ in cognitive ability to complete survey questions. This can influence the eventual results (Gathercole & Alloway 2013; Krosnick, 1991).

Until now, a lot of research has been done to gain more insight into the problems that might arise while completing self-report questionnaires (Galesic et al., 2008; Graesser et al., 2006; Lenzner et al., 2011). In the past, researchers made use of cognitive interviewing techniques such as think-aloud protocols and verbal probing to get a grip on the difficulties that might arise (Collins, 2003). The think-aloud protocol is a data-gathering method in which respondents are asked to verbalise their thought processes during or after doing a specific task. Verbal probing is a cognitive interviewing technique where questions are designed to elicit specific information that is usually not provided by respondents. These cognitive interviews provide a suitable methodology for examining the extent to which tools of inquiry capture the experiences of students in a valid and reliable manner (Beatty & Willis, 2007; Desimone & Le Floch, 2004; Presser et al., 2004). However, despite the benefits cognitive interviews have to offer, they do not allow researchers to look directly into processing behaviour while respondents complete in the questionnaire.

2.2. Eye tracking as an eye-opener in survey research

Previous research has shown that eye tracking can help gain more insight into the black box of the processes of completing self-report questionnaires (Galesic et al., 2008; Redline & Lankford, 2001). Via this relatively unobtrusive instrument, one can track the implicit processes at play while completing questionnaires. Eye tracking research has a long tradition in studying cognitive processing during reading and other information processing tasks (Duchowski, 2007; Neuert, 2016; Rayner, 1998). More recently, the technique has also been introduced into the field of survey methodological research to study cognitive processes while answering survey questions (Lenzner et al., 2010; Neuert, 2016).

In previous research, eye tracking has been used to study, among other topics, the visual designs of branching instructions (Redline & Lankford, 2001), different response formats (Lenzner et al., 2014), response order effects (Galesic et al., 2008), the effects of question wording (Graesser et al., 2006; Lenzner et al., 2011) and the cognitive processes associated with answering rating scale questions (Menold et al., 2014). However, these were mainly experimental studies that focused on the aspects of the questionnaire that could lead to difficulties in processing. By investigating the potential burden the questions might bring, one focuses on the possible constraints of the survey. However, the effects these difficulties have on the actual process of completing the questionnaire have not yet been addressed.

The relationship between eye movements and cognitive processing is based on two assumptions: the immediacy assumption and the eye-mind assumption. The immediacy assumption states that a visual stimulus on which the eyes fixate is processed immediately. The eye-mind assumption postulates that as long as the stimulus is fixated, it is mentally processed. Thus, both assumptions suggest that eye

movements provide direct information about what is processed and the amount of cognitive effort that is involved (Just & Carpenter, 1980).

Although eye tracking cannot help in making a concrete distinction between the different stages respondents might go through when completing self-report questionnaires, it does provide insight into the entire process that evolves in the time period between the reading of the stimulus — in this case, the self-report question — and the giving of an answer. The duration a respondent spends processing gives an indication of the cognitive effort the respondent put into the processing. Longer fixation times could, for example, be associated with a deeper and more effortful cognitive processing or may be an indicator of comprehension problems (Holmqvist et al., 2011).

2.3. Present research

In this study, we will include eye tracking as an online measure in order to gain an understanding of the cognitive processes that are active while completing self-report questionnaires. By taking a closer look at eye tracking data, we strive to examine whether the underlying processes are possibly explanatory indicators of the internal consistency by which the respondent completed the questionnaire. After all, internal consistency is one of the most critical prerequisites in obtaining meaningful results from survey data. Consistency is determined by how similar a respondent answers questions that belong to the same scale. Our study aims to answer the following two research questions:

1. To what extent is there a relation between the consistency in answering behaviour and eye movement measures when completing a self-report questionnaire?
2. To what extent is there a relationship between the consistency in answering behaviour, eye movement measurers and a respondent's working memory capacity when completing a self-report questionnaire?

In the assumption that the time a respondent spends fixating on an area of the survey item more or less corresponds to the time this area is processed (Staub & Rayner, 2007), the time taken to choose an answering option can be an indicator of the cognitive effort that was invested in arriving at this answer or judgment (Fazio, 1990). Therefore, we hypothesise that there could be a link between the cognitive processing taking place when scoring the items of a self-report questionnaire and the internal consistency of the scales. Based on the previous findings on working memory capacity (Krosnick, 1991), we expect an interplay between students' working memory capacity and the cognitive process taking place when completing a questionnaire.

3. Methodology

3.1. Participants

The sample consisted of 92 bachelor students from a social science faculty. Students were recruited during regular lectures and all participated on a voluntary basis. Before the start of the experiment, we received their consent, which was approved by the ethics committee for social sciences and humanities of the participating university. All participants had a normal or corrected-to-normal vision and had Dutch as their native language. Due to issues that are common in eye tracking research (e.g. problems with the calibration of the eye tracker, and a lack of responses to the survey questions [see e.g. Holmqvist et al., 2011]) we lost data from 22 respondents. Data from 10 respondents were excluded due to technical issues; data from 12 respondents were left out because of poor quality of the eye tracking data. After this data cleaning, the data from 70 participants were included in the statistical analyses. To thank the students for their participation, they received two cinema tickets.

3.2. Materials and procedure

The self-report questionnaire data were collected as part of a larger project about learning from texts and the completing of questionnaires where we recorded eye movements to gain insight into the processing behaviour of participants. After the reading of each text, a validated self-report questionnaire was completed to measure students' task-specific processing strategies. A task-specific version of the ILS-SV questionnaire was developed based on the original version (Donche & Van Petegem, 2008; Vermunt & Donche, 2017). This version contained four scales about cognitive processing strategies and consisted of sixteen items that mapped how participants process information when reading a particular text. Students had to read the question, select the answering category of their choice, and state their answer out loud. Answering options ranged from 1 = 'I rarely or never do this' to 5 = 'I almost always do this'. All survey items were answered consecutively without the possibility of changing the given answer.

Apart from completing the self-report questionnaire, the students' working memory capacity was measured by means of the Automated Operation Span Task (Aospan). According to Unsworth et al. (2005), the Aospan is a reliable and valid test for measuring the working memory capacity that can be used in various research domains. Participants were required to solve a series of mathematical operations while trying to retain a set of unrelated letters. The Aospan is mouse-driven, calculates scores automatically and requires little to no intervention from the experimenter (Unsworth et al., 2005). In order to be sure that participants were not only focusing on remembering the letters, a 85% accuracy criterion was imposed for solving the mathematical problems (Unsworth et al., 2005). The Aospan provides two scorings, an absolute credit scoring and a partial credit scoring. Since partial credit scoring is preferred over the absolute all-or-nothing scoring, we made use of the latter (Conway et al., 2005). The mean score for all respondents was 60.11 (SD = 9.85). The score for this working memory capacity test was normally distributed, and for further analysis, we made use of standardised scores.

3.3. Eye tracking equipment

To measure students' eye movements, we made use of the Tobii Pro X3-120 eye tracker, which alternates between bright and dark pupil eye tracking in a predefined, systematic way. This eye tracker had a sampling frequency of 120 Hz (binocularly), which made it possible to take a closer look at the fixation durations. The eye tracker was secured to a 17.3-inch monitor with a resolution of 1.920 x 1.080 pixels. Every participant sat at about 60 cm from the screen and the eye tracker. To minimise the influence of student movement, we employed a chinrest. Tobii Technology (Stockholm, Sweden) reported a gaze accuracy of 0.4°, gaze precision of 0.24° and a total system latency of fewer than 11 milliseconds for this eye tracker. The eye movements were recorded with Tobii-Studio (3.4.8) software.

3.4. Consistency in response behaviour

The first indication of consistency in response behaviour is the Cronbach Alpha coefficient, which was calculated for each of the four scales. The consistency levels for the four scales were .67 .68, .69 and .65, respectively (Table 1). These results show an acceptable internal consistency level for four-item scales. Since only a small number of items are used per scale, and given the sensitivity of the Cronbach's Alpha for the number of items, a cut-off value of .60 is considered sufficient (Cortina, 1993; Pallant, 2007). As respondents can differ in the way they score the separate items of a specific scale, thus showing diversity in scoring behaviour across items, we categorised their rating behaviour for each scale. For all respondents, four consistency indicators were created (one for each scale), making distinctions between raters using the same answering category for all items on a scale or raters showing more diversity in the use of answering categories, by making, for instance, use of at least two different answering categories. The consistency indicator ranged from 1 (consistent answering pattern) to 4 (very diverse answering pattern). The questionnaire did not include reversed items, so this could not serve as an explanation for diversity in answering categories.

Table 1

ILS-SV scales, number of items, item examples and reliability (internal consistency)

Scale	Items	Item example	Cronbach's Alpha
Relating and structuring	4	I compare conclusions from different teaching modules with each other.	.67
Critical processing	4	I try to understand the interpretations of experts in a critical way.	.68
Analysing	4	I study each course book chapter point by point and look into each piece separately.	.69
Memorising	4	I learn definitions by heart and as literally as possible.	.65

3.5. Analysis

We used the Tobii fixation filter for fixation identification, which is an implementation of a classification algorithm proposed by Olsson (2007). It uses a velocity threshold (35 pixels per window) and a distance threshold (35 pixels) (Olsen, 2012).

Eye movement data were analysed at the item level. The question and response field for each item in the survey were considered as a combined area of interest (AOI). For each AOI or item in the survey, the total fixation duration and the total fixation count were calculated separately. To control for the length of AOI's, the total fixation duration measure was normalised by calculating a milliseconds-per-character measure (Ariasi et al., 2017; Catrysse et al., 2016; Yeari et al., 2016). The total fixation count measure was normalised by calculating a count-per-character measure. In addition, we logarithmically transformed these measures because they are heavily skewed (Catrysse et al., 2018; Holmqvist et al., 2011; Lo & Andrews, 2015). To check the distribution of the dependent measures, the fitdistrplus package was used (Delignette-Muller & Dutang, 2015). The eye movement data were analysed with linear mixed effects models (LMM) with the lme4 package (Bates et al., 2015) in R (R Core Team, 2014) and with the Rstudio interface. Mixed-effects models are statistical models that incorporate random and fixed effects (Baayen, 2008; Baayen et al., 2008). Subjects, items and scales were considered as crossed random effects (Baayen, 2008; Baayen et al., 2008). The analysis was conducted at the item level and was based on 1,120 data points (70 students each processing 16 items).

Separate models were fitted for the total fixation duration and the total fixation count. Two models per measure were fitted: (1) an LMM with subjects, subscales and items as random effects and consistency in answering behaviour as a fixed effect and (2) an LMM with subjects, subscales and items as random effects and consistency in answering behaviour and working memory capacity as fixed effects. The interactions between the fixed effects were also incorporated into the second model.

4. Results

4.1. The relation between consistency in answering behaviour and eye movement measures

In order to answer the first research question, we report the means and standard deviations for the eye movement measures in Table 2 in relation to the consistency in answering behaviour. For example, students who were very consistent in their answering behaviour on a certain scale, that is, choosing the same answering option for each item, looked on average 8.74 seconds at an item and the

corresponding answering options, and made on average 31.99 fixations on an item and answering options.

Table 2

Descriptive statistics for the number of different answering options per scale in relation to the eye movement measures

	Consistency level indicator 1		Consistency level indicator 2		Consistency level indicator 3		Consistency level indicator 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total fixation duration (seconds)	8.74	4.56	9.02	4.70	8.95	5.21	8.75	5.80
Total fixation count	31.99	16.30	34.15	17.65	34.49	19.99	32.97	22.72

Note: Untransformed eye movement measures reported.

In the next step, we examined the relation between the consistency in answering behaviour and eye movement measures. We analysed the data with linear mixed effect models. For the total fixation duration, the parameter estimates indicated that there was a significant effect of consistency in answering behaviour on the total fixation duration for an item (Table 3). More specifically, the results showed that a student who chose two or three different answering categories looked longer at an item than a student who only opted for one answering category. A student who chose four categories did not look longer at the items than a student who chose only one answering category.

Overall, the parameter estimates showed that students with less consistent scoring behaviour spend more time on processing the items and answering options. However, this was not the case for students who picked four different answering options on a scale. This implies that there seems to be a turning point in the effect of correlation between consistency in answering behaviour and students' eye movement measures.

Table 3

Parameter estimates of the random and fixed effects for the random intercept model for total fixation duration and total fixation count

	Total fixation duration				Total fixation count			
Random effects	Variance	*SD*			Variance	*SD*		
Subject	.05	.23			.05	.23		
Item	.04	.20			.03	.18		
Subscale	.00	.001			.00	.00		
Residual	.09	.30			.10	.32		
Fixed effects	*β*	*SE*	*t*	*pr(>|t|)*	*β*	*SE*	*t*	*pr(>|t|)*
Intercept	4.68	.06	-69.53	**<.001**	-.90	.07	-13.88	**<.001**
Consistency level indicator 2	.09	.04	2.17	**.03**	.10	.04	2.39	**.02**
Consistency level indicator 3	.11	.04	2.68	**.007**	.12	.04	2.75	**.006**
Consistency level indicator 4	.09	.06	1.57	.12	.10	.06	1.60	.11

Note: Significant values are in bold.

For the total fixation count, the estimate of the intercept had a negative value of -0.90. This is due to the log transformation of the count-per-character measure, which causes small values (<1) to turn into negative values. Moreover, we were mainly interested in the potential change in the fixation count, rather than in its absolute value. Therefore, this negative value was not problematic for the interpretation of our results.

The results for the fixation count are similar as for the total fixation duration. A student who chose two or three answering categories made more fixations on an item than a student opting for only one answering category. A student who chose four categories did not make more fixations on the items than a student who picked only one answering category.

4.2. The relation between consistency in answering behaviour, working memory capacity and eye movement measures

To answer the second research question on the relationship between consistency in answering behaviour, working memory capacity and eye movement measures, we updated the mixed effects model of Table 5 and added working memory capacity as a fixed effect in a new model (Table 4). Both for the total fixation duration and total fixation count, we did not find any significant effect of working memory capacity. We can thus conclude that working memory capacity in this study has no interference with students' eye movement measures when completing this self-report questionnaire.

Table 4

Parameter estimates of the random and fixed effects for the random intercept model for total fixation duration and total fixation count including working memory capacity

	Total fixation duration				Total fixation count			
Random effects	Variance	*SD*			Variance	*SD*		
Subject	.05	.23			.05	.23		
Item	.04	.20			.03	.18		
Subscale	.00	.00			.00	.00		
Residual	.09	.30			.10	.32		
Fixed effects	*β*	*SE*	*t*	*pr(>|t|)*	*β*	*SE*	*t*	*pr(>|t|)*
Intercept	4.68	.07	69.24	**<.001**	-.91	.07	-13.89	**<.001**
WMC	.02	.05	.52	.60	.01	.05	.23	.82
2 AO	.09	.04	2.26	**.02**	.10	.04	2.46	**.01**
3 AO	.11	.04	2.76	**.006**	.12	.04	2.81	**.005**
4 AO	.09	.06	1.50	.13	.09	.06	1.48	.14
WMC*2 AO	-.05	.04	-1.14	.25	-.05	.04	-1.12	.26
WMC*3 AO	-.04	.04	-.93	.35	-.04	.04	-.87	.38
WMC*4 AO	-.07	.05	-1.40	.16	-.09	.05	-1.63	.10

Note: AO: Answering option(s) — WMC: working memory capacity — significant values are in bold.

5. Discussion

Although self-report questionnaires are widely used to map students' processing strategies, there still is a lacuna in the knowledge about the processes at play while respondents complete these questionnaires. By gaining insight into these processes, we want to provide evidence for the debate about the often-reported reliability issues of self-report questionnaires (Richardson, 2004, 2013; Veenman, 2011; Veenman & van Hout-Wolters, 2005). In this exploratory study, we used eye tracking in order to unobtrusively track the processes that are at play while completing a self-report questionnaire

on cognitive processing strategies. Previous research mainly focused on cognitive difficulties that might arise when processing the questionnaire, and therefore the questionnaire's potential limitations to accurately grasp respondents' opinions and beliefs (Galesic et al., 2008; Graesser et al., 2006; Lenzner et al., 2014; Menold et al., 2014; Redline & Lankford, 2001). How these difficulties affect the process of completing the questionnaire, and how they influence the questionnaire's reliability, are two questions that have not been addressed before.

Based on previous research stating that the time a respondent spends fixating on a specific area is more or less equal to the time this area is being processed, the processing time is assumed to be a good indicator of the invested cognitive effort (Fazio, 1990; Staub & Rayner, 2007). Therefore, we believe that there could be a link between the cognitive processing taking place when scoring the items of a survey and the internal consistency of the scored scales from a questionnaire. This concept of consistency is important, given the fact that when a respondent's answering behaviour is not consistent, one can thus start questioning the reliability of the survey data. We first examined the relationship between the consistency in answering behaviour and eye movement measures. Our results demonstrate that the consistency in answering behaviour is significantly related to the total fixation duration for an item. The more a respondent's answers differ in one scale, the longer the respondent looks at the items compared to those who only opt for one answering option. However, no significant difference was found between the respondents choosing one response option and the ones opting for four different answers for items belonging to the same scale. Given these results, there seems to be a turning point in the effect of consistency in answering behaviour. Results suggest that too much pondering over a question does not lead directly to a more consistent answering behaviour. On the contrary, when the respondents spend more time processing a question, they might be trying to process the question more thoroughly to come to an appropriate and thus consistent response, but they just do not succeed in doing so.

Secondly, we aimed to gain more insight into the relation between eye movement measures, answering behaviour and the working memory capacity of the respondent. Previous research on the working memory demonstrated that its capacity is limited and that respondents may therefore not give each answering option as much attention as the one they considered initially (Gathercole & Alloway 2013; Krosnick, 1991). Therefore, we hypothesised that we would find less consistent answering behaviour for the students with a lower working memory capacity. However, both for the total fixation duration as well as for the total fixation count, we did not find any significant effect of working memory capacity. This could be because memory distortions do not play a significant role when this self-report questionnaire is being completed immediately after completing the task that the questionnaire referred to.

6. Limitations and directions for future research

Although our findings show that eye tracking is a promising technique to gain more insight into the process of completing self-report questionnaires, we want to emphasise the exploratory nature of this study and point at some limitations and directions for future research.

The process of completing questionnaires is an extremely complex process. Different theoretical models try to distinguish different stages that possibly play a role when a respondent is cognitively processing a question (see e.g. Karabenick et al., 2007; Tourangeau, 1984). In our study, we considered the question as well as the answering options as one area of interest. This choice allows for an indication of the total time taken until one decides and thus completes the process of filling in the item. More specifically, by focussing on the survey item in its entirety, we took all stages of the different theoretical models into account. In future research, it would be interesting to separate this area into two distinct areas of interest — the question and the answering options — in order to investigate which possible influence each of these areas has on the internal consistency. This would also allow us to further separate the different stages of the theoretical models. However, as these stages do not follow a linear path, separating into different areas of interest will lead to a loss of information. When analysing the question, we could, for example, consider whether different reading processes lead to different outcomes in internal consistency. Hereto, it would also be important to take other eye tracking measures into account.

In our study, we made use of the total fixation duration and the fixation count to map the whole process. However, analysing merely the question would allow us to use other measures such as first pass fixations and second pass fixations (Hyönä et al., 2003; Jarodzka & Brand-Gruwel, 2017) which could possibly shed some light on further difficulties the respondents encountered. Next to looking at the question itself, it could also be clarifying to look at how the respondent processes the different the answering options. The way in which a respondent ponders over a question — merely focusing on one answering option or considering each of the five possibilities — could potentially elucidate their answering behaviour.

Another constraint of this study is that no use was made of complementary data. The use of merely eye tracking data might not provide us with the necessary insight into the reasons why students who respond in a less consistent way take more time to respond to the items, whereas using a multi-method approach to look at the data could help us put different pieces of the puzzle together (Catrysse et al., 2018). As we already know from previous research, a longer reading time can be an indication of several different cognitive processes such as (1) high-level or deeper cognitive processing (Ariasi & Mason, 2011; Holmqvist et al., 2011; Penttinen et al., 2013), (2) strategic attempts to resolve comprehension problems or further text comprehension (Ariasi et al., 2017; Hyönä & Lorch, 2004; Hyönä et al., 2002; Hyönä et al., 2003; Kinnunen & Vauras, 1995), (3) comprehension monitoring (van Gog & Jarodzka, 2013), (4) difficulty with text passages (Rayner et al., 2006) and (5) attempts to reinstate information into working memory in order to elaborate or rehearse that information (Hyönä & Lorch, 2004). However, further research is needed to investigate whether the current insight in the field of text reading also hold for the process of completing survey questionnaires.

A last observation is that when completing the questionnaire, respondents were asked to state their given answer out loud after every question. Knowing that researchers were monitoring their answers could possibly have had an influence on the natural process of completing the questionnaire. For future research, it would therefore be necessary to look at the processes that are at play without verifying for the responses given by respondents. Moreover, it was impossible for the respondents to change their answer on certain questions. Once they provided an answer, the next question was immediately projected without an opportunity for the respondent to change their mind. Considering this in future research, one will be able to search for doubts and changes in the answering process.

7. Conclusions

Notwithstanding certain limitations, our exploratory study was able to show that eye tracking offers important research perspectives that helped us gain more insight into the cognitive processes at play in the process of completing a self-report questionnaire. It also gave us insight into how these processes are related to the consistency by which the survey has been completed. By lifting a corner of the veil that lies over survey research, we now not only know that a longer processing time is not necessarily linked to more consistent answering behaviour, but also that there is a turning point in which longer processing does not lead to more consistency in answering behaviour.

Key points

- The use of eye tracking to record the process of completing self-report questionnaires appears to be a promising tool to gain more insight herein.
- Respondents who look longer at the item in question do not necessarily have more consistent answering behaviour than respondents who spend less time answering the questions.
- When answering self-report questionnaires, there seems to be a turning point in which a longer focus on the item does not lead to a more consistent answering pattern.

References

Ariasi, N., Hyönä, J., Kaakinen, J., & Mason, L. (2017). An eye-movement analysis of the refutation effect in reading science text. *Journal of Computer Assisted Learning, 33*(3), 202-221. https://doi.org/10.1111/jcal.12151

Ariasi, N., & Mason, L. (2011). Uncovering the effect of text structure in learning from a science text: An eye-tracking study. *Instructional science, 39*(5), 581-601. https://doi.org/10.1007/s11251-010-9142-5

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47-89): Academic Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Beatty, P., & Willis, G. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly, 71*(2), 287-311. https://doi.org/10.1093/poq/nfm006

Catrysse, L., Gijbels, D., & Donche, V. (2018). It is not only about the depth of processing: What if eye am not interested in the text? *Learning and Instruction, 58*, 284-294. https://doi.org/10.1016/j.learninstruc.2018.07.009

Catrysse, L., Gijbels, D., Donche, V., De Maeyer, S., Van den Bossche, P., & Gommers, L. (2016). Mapping processing strategies in learning from expository text: an exploratory eye tracking study followed by a cued recall. *Frontline learning research, 4*(1), 1-16. https://doi.org/10.14786/flr.v4i1.192

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of life research, 12*(3), 229-238. https://doi.org/10.1023/A:1023254226592

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769-786. https://doi.org/10.3758/bf03196772

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software, 64*(4), 1-34. https://doi.org/10.18637/jss.v064.i04

Desimone, L., & Le Floch, K. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational evaluation policy analysis, 26*(1), 1-22. https://doi.org/10.3102/01623737026001001

Dinsmore, D., & Alexander, P. (2012). A Critical Discussion of Deep and Surface Processing: What It Means, How It Is Measured, the Role of Context, and Model Specification. *Educational psychology review, 24*(4), 499-567. https://doi.org/10.1007/s10648-012-9198-7

Donche, V., & Van Petegem, P. (2008). The validity and reliability of the short inventory of learning patterns. In E. Cools, H. van den Broeck, & T. Redmond (Eds.), *Style and cultural differences: how can organisations, regions and countries take advantage of style differences* (pp. 49-59). Ghent: Vlerick Leuven Ghent Management School.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice*. London: Springer.

Fazio, R. (1990). Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 75-109). New York: Academic Press.

Fowler, F. (2014). *Survey research methods - 5th edition*. Thousand Oaks: Sage publications.

Galesic, M., Tourangeau, R., Couper, M., & Conrad, F. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72*(5), 892-913. https://doi.org/10.1093/poq/nfn059

Gathercole, S., & Alloway , T. (2013). *De invloed van het werkgeheugen op het leren: Handelingsgerichte adviezen voor het basisonderwijs*. Amsterdam: SWP, Amsterdam.

Graesser, A., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID) a web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3-22. https://doi.org/10.1093/poq/nfj012

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Hyönä, J., & Lorch, R. (2004). Effects of topic headings on text processing: Evidence from adult readers' eye fixation patterns. *Learning and Instruction, 14*(2), 131-152. https://doi.org/10.1016/j.learninstruc.2004.01.001

Hyönä, J., Lorch, R., & Kaakinen, J. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology, 94*(1), 44-55. https://doi.org/10.1037//0022-0663.94.1.44

Hyönä, J., Lorch, R., & Rinck, M. (2003). Eye Movement Measures to Study Global Text Processing. In R. Hyönä (Ed.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 313-334). Amsterdam: Elsevier Science.

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: towards a model of real-world reading. *Journal of Computer Assisted Learning, 33*(3), 193-201. https://doi.org/10.1111/jcal.12189

Jobe, J., & Herrmann, D. (1996). Implications of models of survey cognition for memory theory. *Basic applied memory research, 2*, 193-205. https://doi.org/10.1023/A:1023279029852

Just, M., & Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review, 87*(4), 329. https://doi.org/10.1037/0033-295X.87.4.329

Karabenick, S., Woolley, M., Friedel, J., Ammon, B., Blazevski, J., Bonney, C., . . . Kempler, T. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*(3), 139-151. https://doi.org/10.1080/00461520701416231

Kinnunen, R., & Vauras, M. (1995). Comprehension monitoring and the level of comprehension in high- and low-achieving primary school children's reading. *Learning and Instruction, 5*(2), 143-165. https://doi.org/10.1016/0959-4752(95)00009-R

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Krosnick, J., & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219. https://doi.org/10.1086/269029

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23*(3), 361-373. https://doi.org/10.1093/ijpor/edq053

Lenzner, T., Kaczmirek, L., & Galesic, M. (2014). Left Feels Right: A Usability Study on the Position of Answer Boxes in Web Surveys. *32*(6), 743-764. https://doi.org/10.1177/0894439313517532

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003-1020. https://doi.org/10.1002/acp.1602

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using Generalized Linear Mixed Models to analyse reaction time data. *Frontiers in Psychology, 6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Marsden, P., & Wright, J. (2010). *Handbook of survey research - 2nd edition*. Bingley: Emerald Group Publishing.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*(1), 21-39. https://doi.org/10.1177/1525822X13508270

Neuert, C. (2016). *Eye tracking in questionnaire pretesting.*

Olsen, A. (2012). The Tobii I-VT fixation filter.

Olsson, P. (2007). Real-time and offline filters for eye tracking.

Pallant, J. (2007). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows - 3th edition*: Maidenhead: Open University Press.

Penttinen, M., Anto, E., & Mikkilä-Erdmann, M. (2013). Conceptual change, text comprehension and eye movements during reading. *Research in Science Education, 43*(4), 1407-1434. https://doi.org/10.1007/s11165-012-9313-2

Presser, S., Couper, M., Lessler, J., Martin, E., Martin, J., Rothgeb, J., & Singer, E. (2004). Methods for Testing and Evaluating Survey Questions. *Public Opinion Quarterly, 68*(1), 109-130. https://doi.org/10.1093/poq/nfh008

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Chace, K., Slattery, T., & Ashby, J. (2006). Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading, 10*(3), 241-255. https://doi.org/10.1207/s1532799xssr1003_3

Redline, C. D., & Lankford, C. (2001). Eye-movement analysis: a new tool for evaluating the design of visually administered instruments (paper and web). *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Richardson, J. (2004). Methodological Issues in Questionnaire-Based Research on Student Learning in Higher Education. *Educational psychology review, 16*(4), 347-358. https://doi.org/10.1007/s10648-004-0004-z

Richardson, J. (2013). Research issues in evaluating learning pattern development in higher education. *Studies in Educational Evaluation, 39*(1), 66-70. https://doi.org/10.1016/j.stueduc.2012.11.003

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 40 - 62
ISSN 2295-3159

Opening the black box of students' text-learning processes: A process mining perspective

Amelie Rogiers[a], Emmelien Merchie[a] & Hilde van Keer[a]

[a]Department of Educational Studies, Ghent University, Ghent, Belgium

Article received 26 June 2019 / Article revised 11 September / Accepted 3 January / Available online 30 March

Abstract

The current study uncovers secondary school students' actual use of text-learning strategies during an individual learning task by means of a concurrent self-reported thinking aloud procedure. Think-aloud data of 51 participants with different learning strategy profiles, distinguished based on a retrospective self-report questionnaire (i.e., 15 integrated strategy users, 15 information organizers, 10 mental learners, and 11 limited strategy users), were analysed by means of educational process mining. Both the frequency of students' strategy use, as well as the temporal patterns between these strategies were studied. The process mining results clearly demonstrated differences between the strategy profiles with respect to the frequency of their applied strategies, as well as concerning the temporal sequences wherein strategies were applied throughout the course of students' text-learning process. The added value of combining both retrospective and concurrent self-report measures of students' strategies as well as conducting process mining analysis is discussed.

Process mining; learner profiles; think-aloud protocol analysis; on-line measures; off-line measures

Corresponding author: Amélie Rogiers, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium, E-mail: Amelie.Rogiers@UGent.be *DOI: https://doi.org/10.14786/flr.v8i3.527*

1. Introduction

Recently, both educational researchers and practitioners have emphasized the importance of adjusted or personalized curricula wherein both the instructional content and methods are tailored to students' individual learning needs (Deed et al., 2014). This is also recognized by the OECD Learning Framework 2030 (2018) advocating the importance of learner-oriented teaching and learning. In view of contributing to the evidence-based design of personalized curricula, educational researchers are concerned with both measures and data analysis approaches to fully map and understand individual students' learning. Considering these *measures*, it is clear that the inclusion of both off-line and on-line instruments for measuring students' learning is preferable given their complementary properties (Veenman, 2011). While off-line measures are administered prospectively or retrospectively to performance on a learning task (e.g., self-report questionnaire data), on-line measures are gathered concurrently during task performance (e.g., think-aloud protocol or verbal self-report data). Consequently, while off-line measures enable researchers to uncover learners' perceptions of which and how often certain strategies are applied during learning, on-line measures additionally enable to map how and when these strategies are actually applied throughout the learning process (i.e., in which sequence strategies are applied or which switches occur between strategies; Merchie & Van Keer, 2014). In this respect, researchers increasingly advocate to combine both measures in view of gaining rich and detailed insight into both students' perceptions and actual strategic behaviour (Bråten & Samuelstuen, 2007; Veenman, 2005).

As to the *data analysis approaches* for gaining insight into students' learning processes, researchers call progressively for applying a more person-oriented approach, next to the rather dominant variable-oriented approach focusing primarily on analysing relationships among variables (Alexander et al., 2018; Fryer & Vermunt, 2017). Such a person-oriented approach is highly recommended as it emphasizes the study of naturally occurring clusters or profiles in students' learning (Bergman et al., 2003). Stemming from a person-oriented approach on students' text-learning strategies (i.e., strategies to select, organize, condense, and retain text information in a more memorable form; Rogiers et al., 2019a; Weinstein et al., 2011), previous research already succeeded to identify learning strategy profiles in a large sample of 1,931 secondary school students (Rogiers et al., 2019a). Four learning strategy profiles, in which students differently combine diverse strategies during text learning, were determined based on a retrospective self-report questionnaire. More particularly, *integrated strategy users* (ISU) were identified as learners with the most preferable profile, as they engaged in the strategic combination of different covert (i.e., non-observable, e.g., elaborating) and overt (i.e., observable, e.g., summarizing), cognitive (e.g., elaborating), and metacognitive (e.g., monitoring) text-learning strategies, and outperformed their peers on a subsequent performance test. The *information organizers* (IO) frequently applied text-noting strategies (i.e., highlighting, summarizing) and reported limited use of mental learning strategies. Conversely, *mental learners* (ML) restricted their repertoire to covert mental learning strategies (i.e., rereading, paraphrasing) without text-noting strategy use. Finally, *limited strategy users* (LSU) were considered as the non-strategic or less preferable profile, as they mainly focused on the frequent application of one single text-learning strategy (i.e., highlighting, rereading) and obtained the lowest performance scores afterwards. These learning strategy profiles were also identified in late elementary education (Merchie et al., 2014) and in subsequent samples of secondary school students (Rogiers et al., 2018, 2019a). Hence, the abovementioned learning strategy profiles were already corroborated several times in different age groups and independent study samples.

To date, however, there is a gap in the literature when it comes to research providing insight into the temporal sequences in which certain strategies are applied differently by learners during the course of their text-learning process. In this respect, it is seldom investigated which strategy switches unfold during this process (Cromley & Wills, 2016). Although current theories of (text) learning implicitly or explicitly state to account for what happens during this process, this matter has rarely been tested empirically with sequential analyses of real-time process data (e.g., from think-aloud protocol transcripts; Cromley & Wills, 2016). In this respect, the present study contributes to the first explicit question regarding self-report data tackled throughout the different contributions in this special issue. More particularly, it is believed that the learning process of a strategic learner can be characterised as cyclical and adaptive. First, from a self-regulated learning (SRL) perspective, students' learning process is considered as a cyclical process, consisting of different phases occurring before, during, and after learning (i.e., forethought, performance, and reflection phase; Zimmerman, 2002). These phases are not viewed as linearly structured, but considered dynamic and iterative (Panadero, 2017; Pintrich, 2000; Zimmerman, 2002). Second, next to the general comprehensive models of SRL, also more domain-specific learning strategy models (i.e., Good Strategy User Model by Pressley et al., 1987; Model of Strategic Learning by Weinstein et al., 2011; Model of Domain Learning by Alexander, 1998) point to the importance of adaptive strategy use, which encompasses engaging deliberately and flexibly in the use of various strategies. Rather than following a linear and rigid approach to text learning, strategic learners are believed to undertake learning in an adaptive way, wherein

they interactively return to prior learning activities or phases when necessary (Alexander & Jetton, 2000; McNamara et al., 2007; Simpson & Nist, 2000; Wade et al., 1990). Prior research already indicated that high achievers appear to be mostly integrated strategy users, adopting various strategies (Merchie et al., 2014; Rogiers et al., 2019a) and regulate their learning process more effectively (Stoeger et al., 2015). However, it is unclear whether these process statements as put forward in different theoretical models can be grounded empirically and whether

and how exactly specific strategy sequences unfold during the course of learners' text-learning process (Cromley & Wills, 2016). In this respect, it is often difficult to grasp the cyclical and adaptive nature of learning processes as described in the abovementioned theoretical models by means of retrospective self-report measures after learning occurred. It is therefore necessary to analyse students' learning process in a more fine-grained way (i.e., occurrence after occurrence) as it unfolds in real time during learning. Opening this black box and gaining insight into the cyclical and adaptive nature in which particular sequences and strategies unfold throughout students' learning process can offer valuable starting points for providing learner-oriented teaching and learning. If certain effective sequences between strategies come to the fore, for example, then not only strategies, but also effective sequences of applied strategies should be taught (e.g., from one learning strategy to another).

A promising and emerging technique to gain systematic insight into these sequences and analyse students' concurrent self-reports (e.g., think-aloud protocols) is Educational Process Mining (EPM). The idea behind EPM is to discover, monitor, and improve students' actual learning processes by extracting knowledge from recorded time stamps (Bannert et al., 2014). A time stamp refers to the moment wherein the learner is executing or initiating a certain learning activity (e.g., highlighting, rereading). By means of these timestamped activities derived from learners' observed learning behaviour, compact educational process models are composed (Van der Aalst, 2011). These process models provide an overview of both learners' executed activities and the paths that occur between these activities. Whereas the *activities* map the number and frequency of certain applied strategies, the *paths* represent how, and in which sequences these strategies were adopted throughout the learning process (Fluxicon, 2019; Van der Aalst, 2011). As such, EPM enables to visualise students' learning behaviour and facilitates a thorough understanding of the course of students' complex real-time learning process. In the context of SRL, for example, EPM research has shown that university students' sequences of self- or group-regulatory activities differed among successful and less successful students (e.g., Schoor & Bannert, 2012). However, the application of EPM in educational research is still in its infancy and, to date, the temporal order of students' applied strategies during task completion has been widely neglected (Bannert et al., 2014; Reimann, 2007). More in-depth EPM analyses can, therefore, yield valuable insights into students' learning process and can complement off-line measures of students' applied strategy use. In this respect, it also enables to investigate to which degree retrospective self-report measures accurately reflect students' actual strategy use that is revealed while concurrently thinking aloud. Our study adds to the literature by systematically analysing real-time think-aloud protocol (further referred to as 'TAP') data from students with different learning strategy profiles who are requested to learn an informative text and by considering strategy sequences by means of EPM. Further, this study adds to the literature by confronting the frequency of students' text-learning strategies as measured via concurrent measures on the one hand (i.e., TAP) and retrospective measures on the other hand (i.e., a task-specific self-report questionnaire) and study their overlap (Rogiers et al., 2019b).

1.2 The present study

By means of EPM, the current study investigates students' actual use of text-learning strategies when executing an independent learning task while thinking aloud. In a first step, this study aims to examine the *frequency* of students' occurred text-learning strategies depending on their learning strategy profile (RQ1). Referring to previous research using task-specific self-report questionnaires (Merchie et al., 2014; Rogiers et al., 2018, 2019a), we hypothesize more frequent verbalisations of various text-learning strategies in integrated strategy users, and less frequent and diverse strategy verbalisations in limited strategy users. In addition, we expect more frequent verbalizations of the application of overt text-noting strategies (e.g., summarizing) in information organizers and a predominant application of covert mental learning strategies (e.g., paraphrasing) in mental learners. In a second step, this study aims to explore *temporal patterns* in students' text-learning process based on the sequences in which their strategies are applied (RQ2). As the theoretical and empirical literature indicates that particularly effective learners apply diverse strategies in a flexible and systematic way (Alexander & Jetton, 2000; McNamara et al., 2007; Rogiers et al., 2019a; Simpson & Nist, 2000; Wade et al., 1990; Weinstein et al., 2011), we hypothesize a more cyclical use of text-learning strategies in integrated strategy users, including more recursive

patterns between their applied strategies. Conversely, a more linear and unidirectional text-learning process is expected in limited strategy users.

2. Methodology

2.1 Participants

A think-aloud study was carried out with 51 secondary school students (62.75% seventh and 37.25% eight graders) from 11 schools and 51 classes who were part of a large-scale study (n = 1,931, Rogiers et al., 2019a). Based on a large-scale cluster analysis, 15 integrated strategy users, 15 information organizers, 10 mental learners, and 11 limited strategy users (n = 51 participants) were identified within the sample of the think-aloud study. The sample consisted of 70.59% girls and 29.41% boys, with an overall mean age of 12.99 years (SD = .69). The majority of the students (87.23%) were native Dutch speakers, which is the language of instruction in Flanders (the Dutch speaking part of Belgium). All participants and their parents agreed to participate in the TAP administration by means of informed consent.

2.2 Instruments and procedure

The data collection procedure consisted of several steps. Figure 1 provides a visual representation of the data collection procedure.

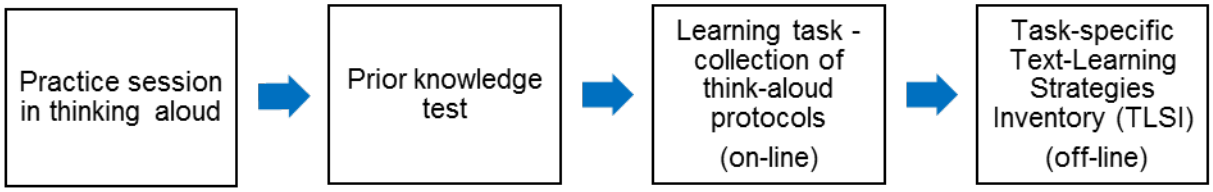


Figure 1. Chronological representation of the data collection procedure.

As can be seen in both last steps of the procedure, a combination of two self-report measures was opted for in the context of the present study, respectively an on-line and concurrent think-aloud measure on the one hand and an off-line, retrospective questionnaire on the other hand.

2.2.1 Practice session in thinking aloud

Following the recommendations of prior research (Greene et al., 2011; van Someren et al., 1994), a 20-minute practice session in thinking out loud was organised by the researcher to familiarize students with the think-aloud method. This practice session was based on prior research in a similar age group (Merchie & Van Keer, 2014; Vandevelde et al., 2005) and consisted of three phases. In a first phase, the researcher thoroughly explained the purpose and procedure of the think-aloud method. Second, the researcher modelled thinking aloud during an origami assignment. No learning task was opted for practicing thinking out loud to avoid possible training effects (Afflerbach & Johnston, 1984; Greene et al., 2011). The origami assignment provided ample opportunities for self-regulation. For instance, a step-by-step approach could be followed and there were ample opportunities for students to evaluate or adjust their approach. Third, an individual practice phase took place in which students practiced thinking out loud. During this session, students were asked to fold an origami cat while verbalizing their thoughts, feelings, and actions. During the practice session, feedback was provided on students' verbalisations in view of optimizing their thinking aloud. Accordingly, no feedback on students' approach was provided. The researcher prompted the student to continue verbalizing when (a) meaningful silences or (b) certain nonverbal behaviours took place (i.e., frowning, repeatedly turning the text page, staring; Merchie & Van Keer, 2014; Vandevelde et al., 2015), thereby thoughtfully considering the student and situation at hand to avoid the loss of meaningful information about students' behaviour (e.g., Boekaerts & Corno, 2005). As prompt, students were consistently given the instruction: "verbalize everything that you are doing or thinking" or "keep thinking aloud". In this respect, type 1 (verbal content) and type 2 (nonverbal content) verbalizations were encouraged, and type 3 verbalizations were avoided since students were not asked to explain their cognition. Consequently, researchers

were able to identify spontaneous self-regulatory learning activities (Ericsson & Simon, 1980; Vandevelde et al., 2015).

2.2.2 Prior knowledge test

As prior knowledge might influence text learning (Alexander & Jetton, 2000; Bråten & Samuelstuen, 2004), a prior knowledge test regarding the text topic was administered before the actual learning task. Students were asked to write down everything they already knew about the topic. Following prior research, the matching of students' notes to the text content was opted for to score the prior knowledge test (for more information on this procedure, see Merchie et al., 2014) The matching of students' notes to the text content revealed very limited to no prior knowledge regarding the text content (M = 5.66, SD = 2.71; Min = 0, Max = 24).

2.2.3 Learning task

Since studying in preparation for a classroom test is a regular task in secondary education, students were instructed to study an informative text in the way they would prepare for a test while thinking out loud (Fox, 2009; Samuelstuen & Bråten, 2007). For the learning task, a 442-word informative text was used of which the participants did not study the topic (i.e., chewing gum) as part of their courses. The multi-paragraph text consisted of one title (i.e., chewing gum), four sections and subtitles (i.e., history, production, advantages, and disadvantages), and three pictures. Text quality was verified in advance (see Rogiers et al., 2019a). In view of encouraging students to plan their work, they were informed to have 50 minutes time for task completion. To enable students to monitor their progress, a clock was provided, but no further time indications were given to prevent the prompted monitoring of time. In line with previous studies (Slotte et al., 2001), students were allowed, but not obligated to make notes on scratch paper while studying. During the task completion process, students were observed by the researcher and were only prompted to continue verbalizing when necessary (Greene et al., 2011).

2.2.4 Task-specific self-report inventory

Immediately after learning task execution, students completed the 'Text-Learning Strategies Inventory' (TLSI; Merchie et al., 2014). This task-specific questionnaire consists of 37 items, subdivided into nine subscales (see Appendix A) to which students respond on a five-point Likert-scale (1 = *completely disagree*, 5 = *completely agree*). In line with theoretical frameworks (Wade et al., 1990; Zimmerman, 2002), the TLSI incorporates both cognitive (e.g., paraphrasing) and metacognitive (e.g., monitoring) text-learning strategies, as well as overt (e.g., summarizing) and covert (e.g., paraphrasing) strategies. Good model fit results were obtained for this nine-factor model in prior large-scale research (Rogiers et al., 2019a). Appendix A presents the descriptive statistics and reliability coefficients for the TLSI-subscales. By means of hierarchical and K-means cluster analyses on the TLSI-subscale scores within the larger sample (n = 1,931), students learning strategy profile was determined (for a detailed description, see Rogiers et al., 2019a).

2.3 Think-aloud coding procedure of learning strategies

In a first step, think-aloud sessions were transcribed and coded. As all sessions were audio- and videotaped, both students' verbal and non-verbal behaviour (e.g., highlighting text) was transcribed to increase coding accuracy (Veenman, 2011; Young, 2005). Transcriptions were made by means of a computer program for subtitling videos (i.e., Subtitle Workshop 4). This program enables researchers to register the start and end time of each verbalisation and action. This is essential in view of conducting EPM, as the sequence of strategies is calculated based on their exact time frame. Subsequently, transcripts were segmented by one researcher into units of meaning, with one unit referring to a thematically consisted verbalization of a single text-learning activity (Scott, 2008; van Someren et al., 1994). Repeated actions were analysed as separate activities in view of considering the recurrence of different text-learning strategies. As a result, 1,015 minutes of thinking aloud, and 4,107 units of meaning were identified and coded by means of the coding scheme based on prior research of Merchie and Van Keer (2014). This coding scheme is an adapted version of the 'Text-Learning Strategy Protocol' (TLSP; see Table 1), comprising 11 subcategories referring to different text-learning strategies. In line with the self-report questionnaire, the coding scheme reflects both cognitive and metacognitive, as well as overt and covert strategies. Mean learning time was 20 minutes (SD = 3.89), with a minimum of 6 and a maximum of 34 minutes. Analysis of variance showed no statistically significant differences between the four strategy profiles in terms of their mean learning time, $F(3, 50) = 2.437$, $p = .076$). Finally, two trained coders independently double-coded 27% of the protocols, resulting in high interrater reliability (Krippendorff's $\alpha = .95$; Hayes & Krippendorff, 2007).

Table 1

Coding scheme for analysing students' learning activities

TLSP-categories	Description	Example
Overt cognitive strategies		
Summarizing	Noting words or sentences on the scratch paper, making a summary or scheme	*'So, this part is done (student draws a line underneath his summary).'*
Highlighting	Structuring the text or one's own notes	*'I mark all these important words.'*
Covert cognitive strategies		
Memorizing	Memorizing by rereading the text or one's own notes	The student rereads his scratch paper notes.
Initial reading	Reading the text for the first time	The student reads the text for the first time.
Rereading	Rereading the source text	The student rereads the text out loud.
Rehearsing	Rehearsing the text information	*"Now I will rehearse this part again by heart."*
Paraphrasing	Retelling the text content in one's own words	*'So, in sugar-free gum, xylitol prevents tooth decay.'*
Elaborating	Activating or relating prior knowledge to the text content or providing personal remarks regarding the text content	*'Synthetic rubber is indeed petroleum, we learned that in geography!'*
Metacognitive strategies		
Planning	Exploring the text subject, detecting task demands, planning the strategic approach	*'I will first read the text, then underline important words and then try to memorise the text.'*
Monitoring progress	Checking progress during task performance, reflecting on the available time and the time schedule, monitoring the strategic approach	*'So this is what I have already done and this is what I have to do next'.'*
Monitoring comprehension	Detecting lack of comprehension or mistakes, mentioning awareness of understanding	*'Synthetic rubber… I don't know what 'synthetic' means.'*
Other*	Asking questions to the researcher about the overall procedure of the study, phases of silence…	*'Can I write my name on these papers?'*

Note. TLSP = Text-Learning Strategy Protocol. * In accordance to Schoor and Bannert (2012), this category was excluded from the process mining analysis, as we wanted to concentrate on task-related behaviour.

2.4 Process mining analysis on the think-aloud data

In a next step, the coded learning activities of each learner profile were analysed separately via process mining using Disco (Fluxicon, 2019). This software program enables researchers to study the course of students' actual learning processes by generating process models for each learner profile. In these process models, both (1) the *activities* performed by the learners (i.e., the executed strategies during text learning), and (2) the *paths* or connections that occurred between these activities are displayed (Fluxicon, 2019; Van der Aalst, 2011). Thus, whereas the *activities* refer to the extent in which certain text-learning strategies are adopted (i.e., boxes in Figures

2-5), the *paths* visualize the sequence of these performed activities (i.e., arrows in Figures 2-5). Above these paths, the frequency of each of these sequences is represented. Further, both unidirectional paths (→), bidirectional paths (⇆), and loops (↻) are depicted in the process models, indicating that activities have respectively been conducted in consecution, in alternation, or that the same activity was performed several times in succession.

In line with prior research in the field of SRL (Bannert et al., 2014; Schoor & Bannert, 2012) the fuzzy miner algorithm in Disco was used to perform the analysis. This algorithm relies on two metrics (i.e., significance and correlation) to calculate which activities and paths should be included in the process models and which to be excluded (Günther & van der Aalst, 2007). *Significance* refers to the relative importance of activities and paths, implying that more frequent text-learning activities are retained in the model. *Correlation* is deployed for selecting only paths of closely connected activities (Günther & van der Aalst, 2007). By means of this algorithm, Disco automatically includes strategies and paths that are often conducted by a large group of students in the process model, while less frequent activities and paths or paths that have been seldom conducted by only few students are excluded. To date, however, no specific standards are available on the amount of activities and paths that should be included in the process models. Researchers argue that the ideal number of activities and paths strongly depends on the type of research data and questions involved (Fluxicon, 2019). In general, the inclusion of as much activities and paths as possible while simultaneously avoiding too complex process models is recommended (Fluxicon, 2019). In the current study, the percentages of included activities and paths in students' process models were carefully deliberated among four experts on text learning and SRL. In this respect, 33.33% of the most frequent strategies and the 33.33% most frequent connections between these strategies were included in the analysis. As a result, initially coded categories such as paraphrasing and elaborating (see Table 1) were not included in the 33.33% process models (Figures 2-5). Although these strategies occurred commonly in the group of integrated strategy users and mental learners (Table 2), they were adopted by a rather small share of learners compared to the occurrence of the other strategies. Put differently, these activities did not belong to the 33.33% most frequent activities conducted at least once by a large group of learners.

As a final step, following Schoor and Bannert (2012) and in view of obtaining split-half-reliability for the generated process models, we repeated the EPM analyses for the five most typical individuals of each learning strategy profile. We perceived students as typical integrated strategy users (ISU) when high frequencies were found for different text-learning strategies, whereas typical limited strategy users (LSU) were characterized by the dominant application of only one strategy (e.g., highlighting, rereading). For typical information organizers (IO) and mental learners (ML), strategies with high frequency were respectively text-noting strategies (e.g., highlighting, summarizing for IO) and mental learning strategies (e.g., rereading, rehearsing for ML; Rogiers et al., 2019a). The obtained models for the five most typical individuals of each learning strategy profile were very similar to those in Figures 2-5.

3. Results

3.1 Frequency of occurrence of text-learning strategies in different learning strategy profiles' text-learning process (RQ1)

In view of the first research question, we examined which text-learning activities were executed by the different learning strategy profiles during actual text learning. Table 2 displays the frequency of occurrence of all text-learning strategies included in the process models, as well as the number of students conducting each strategy at least once. One-way analysis of variance was used to test differences between the four learning strategy profiles regarding students' use of different strategies. Additionally, Post Hoc pairwise tests with Bonferroni correction were conducted to investigate these differences in-depth. The analysis revealed significant differences between the four learning strategy profiles (see Appendix B for detailed results of the Post Hoc Pairwise comparisons and effect sizes).

As can be derived from Table 2, the results with regard to students' *cognitive strategy use* show that integrated strategy users (ISU) generally executed more diverse text-learning strategies than information organizers (IO), mental learners (ML), and limited strategy users (LSU). The frequency of occurrence of most strategies was higher in integrated strategy users, as well as the percentage of students that adopted the strategies at least once. The most occurring cognitive strategies for integrated strategy users were summarizing, highlighting, paraphrasing, and elaborating, whereas for limited strategy users highlighting, memorizing, and elaborating were

the most frequent strategies. For mental learners, summarizing and highlighting activities seldomly occurred, whereas rereading and rehearsing were frequently coded. In contrast to mental learners, summarizing and highlighting frequently occurred in the information organizers group, in addition to rereading and rehearsing. These results were reflected in the Post Hoc pairwise test results.

As can be derived from Appendix B, a statistically significant difference between the four learning strategy profiles was found for the verbalized overt text-learning strategies: summarizing ($F(3, 4732) = 85.59$, $p < .001$) particularly in favour of integrated strategy users and information organizers, and highlighting ($F(3, 4732) = 43.88$, $p < .001$) in favour of all learning strategy profiles, except for mental learners. Regarding the covert text-learning strategies, the results show that limited strategy users more frequently applied memorizing ($F(3, 4732) = 4.54$, $p = .004$) than information organizers. Further, rereading ($F(3, 4732) = 22.89$, $p < .001$) and rehearsing ($F(3, 4732) = 144.34$, $p < .001$) were most frequently executed by mental learners and information organizers and less frequent by limited strategy users, whereas paraphrasing ($F(3, 4732) = 21.12$, $p < .001$) and elaborating ($F(3, 4732) = 5.252$, $p = .001$) were less frequent performed by information organizers. No statistically significant difference between the four profiles was found regarding the execution of initial reading ($F(3, 4732) = 2.61$, $p = 0.05$).

The results with respect to students' *metacognitive strategy use* reveal no statistically significant differences between the four profiles regarding the use of comprehension monitoring activities ($F(3, 4732) = 1.10$ $p = 0.348$). In contrast, a statistically significant difference between learning strategy profiles was found with regard to planning ($F(3, 4732) = 38.56$, $p < .001$), indicating that planning activities were particularly executed by limited strategy users. In addition, a statistically significant difference with regard to progress monitoring ($F(3, 4732) = 15.57$, $p < .001$) reveals that this strategy frequently occurred in integrated and limited strategy users. However, in order to gain insight into the sequences in which these different strategies are adopted throughout students' learning process, a closer look at the process models is needed (see RQ2).

Table 2

Frequency of occurrence of text-learning strategies for each learner profile (n = 51), including absolute frequency and number of students [a]

Coding category	ISU (*n* = 15)		IO (*n* = 15)		ML (*n* = 10)		LSU (*n* = 11)	
	Absolute frequency	*n*	Absolute frequency	*n*	Absolute frequency	*n*	Absolute frequency	*n*
Overt cognitive strategies								
Summarizing	296	15	268	14	8*	3	64*	4
Highlighting	273	15	245	12	31*	2	224	11
Covert cognitive strategies								
Memorizing	203	15	181	14	143	8	200	10
Initial reading	283	15	242	15	162	10	131	11
Rereading	31*	10	66	12	47	8	199	11
Rehearsing	45*	9	88	12	189	9	20*	6
Paraphrasing	67*	10	34*	8	83*	7	35*	7
Elaborating	38*	11	13*	5	19*	4	32*	6
Metacognitive strategies								
Planning	18	12	88	12	38	9	123	11
Monitoring progress	77	14	28*	9	9*	4	46*	6
Monitoring comprehension	21*	7	31*	8	20*	5	17	7

Note. [a] Number of students conducting the strategy at least once. * Strategies not included in the process models as they do not belong to the 33.33% most frequent activities that are conducted by a large group of students (see method section and RQ2). *ISU* = integrated strategy users, *IO* = information organizers, *ML* = mental learners, *LSU* = limited strategy users.

3.2 Temporal patterns in the different learner profiles' text-learning process

Figures 2-5 display the resulting 33.33% process models for each learning strategy profile. The direction of the arrows represents the order in which the text-learning activities were adopted throughout students' text-learning process. Strategies that took place in the beginning of students' text-learning process (e.g., planning, initial reading) are depicted at the top of the figures, while strategies that were executed at a later moment or at the end of the learning process (e.g., highlighting, summarizing, memorizing, rereading, rehearsing) are represented at respectively the centre or bottom of the figures.

When contrasting the process models of the different learning strategy profiles and focussing on the cognitive and metacognitive strategies that were included in the models, clear differences can be noticed. First, the *cognitive* activities that were included in students' process models indicate that integrated strategy users (ISU), information organizers (IO), and limited strategy users (LSU) applied a combination of both overt (e.g., summarizing and highlighting) and covert strategies (e.g., rehearsing, memorizing, rereading) during text learning, whereas mental learners (ML) exclusively adopted covert strategies. When considering the overt strategies, the process models show that integrated strategy users, information organizers, and limited strategy users frequently applied highlighting, whereas only the models of integrated strategy users and information organizers include summarizing strategies. Regarding the covert strategies, the results show that rehearsing was only included in the process models of information organizers and mental learners. Second, planning was included as a *metacognitive* strategy in all process models, while progress monitoring was only included in integrated strategy users' process model, indicating that – compared to the frequency of the other strategies – a large share of integrated strategy

users frequently tracked and controlled their progress throughout their learning process. The same applies for comprehension monitoring, which was only included in limited strategy users' process model, implying that a large group of limited strategy users actively monitored (a lack of) understanding while processing the text.

When studying the sequences in which these different strategies were adopted throughout students' learning process, differences in the phases of the text-learning process can be identified. First, when focussing on the *beginning* of students' learning process (i.e., start symbol in the process model), the results indicate that integrated and limited strategy users initiated their learning process with planning before initially reading the text. In contrast, information organizers and mental learners immediately started reading the text without planning in advance, which is indicated by the unidirectional arrows between initial reading and planning. Subsequently, they performed planning after they initially read the text.

Concerning the strategies conducted *during* actual text studying, differences between the four learning strategy profiles were found as well. In the group of limited strategy users (Figure 5), the unidirectional arrows between the different strategies indicate that planning, initial reading, highlighting, memorizing and rereading were consecutively executed. In addition, the unidirectional connection between initial reading and comprehension monitoring in limited strategy users' process model denotes that 18% of these strategy users monitored their understanding after reading the text. This strongly differs from integrated strategy users' process model (Figure 2). While highlighting is also preceded here by initial reading, bidirectional paths are found between initial reading and highlighting, as well as between initial reading and summarizing. Yet, the arrows connecting the different cognitive strategies, as well as the presence of reciprocal arrows, indicate that integrated strategy users alternately adopted these strategies before they started to memorize the text. Further, the position of progress monitoring as rather isolated from the other activities in these strategy users' process model must be noticed. This position is due to the fact that progress monitoring was applied before and after a wide variety of activities, suggesting that integrated strategy users tracked and controlled their progress throughout the entire learning process. However, since the process model only represents 33.33% of the performed activities, the wide variety of arrows were omitted by the program. When analysing the process model in detail, the large number of arrows between progress monitoring and a diverse set of text-learning strategies can be found. Furthermore, it is notable that integrated strategy users (Figure 2) alternated strategies before learning (i.e., planning) with activities during learning, which is indicated by the bidirectional arrows connected to students' planning strategy. More particularly, they considered their planning not only before reading in the pre-learning phase, but throughout the different phases in their learning process (i.e., after reading, highlighting, and memorizing).

When taking a closer look at information organizers' process model (Figure 3), many paths are visible, demonstrating that information organizers frequently switched between various strategies throughout their learning process, or frequently resumed previous strategies. This reveals that their text-learning process was rather cyclical organized. Especially the strategies 'summarizing' and 'highlighting' played a prominent role in these learners' learning process, as can be derived from the large number of incoming and outgoing arrows. For instance, after initially reading the text, the unidirectional arrows indicate that information organizers considered their planning before summarizing. After summarizing, a large share of these learners returned to reading the text or started memorizing or rehearsing the text. A clear bi-directional path is present between memorizing and highlighting activities, indicating that these activities were performed in alternation. Further, highlighting was also frequently followed by rehearsing, initial reading, and/or summarizing. Remarkable is that after conducting memorizing, summarizing, and highlighting activities, a large share of information organizers returned to initially reading the text. This could imply that these learners started to engage in different text-learning strategies without first reading or fully understanding the study text.

Although both information organizers' and mental learners' learning process was initiated by initial reading and planning, the further course of their learning process clearly differed. While a large share of information organizers started summarizing the text, a large share of mental learners (Figure 4) started memorizing the text after planning. Further, the recursive loop for memorizing, rehearsing and rereading demonstrates an alternated application of these strategies in mental learners, while the unidirectional arrows show that rehearsing was often followed by rereading and rereading by memorizing.

More fine-grained differences in the course of students' text-learning process can be detected when taking a more detailed look at the direction of the arrows in the process models for each learning strategy profile. The results show that limited strategy users followed a mainly linear structured learning process, as is indicated by the unidirectional arrows in their process model and the absence of any bidirectional paths. In contrast, the other three profiles returned more to prior activities or phases throughout their learning process, which is illustrated by the

returning arrows pointing from the bottom to the top. Hence, no strict linear, but rather a cyclical approach to learning was followed by these profiles.

At last, the arrows leading to 'stop' in the process models (i.e., stop symbol in the figures) show which strategies were conducted *lastly* by the learners. As can be derived from the models, across all learning strategy profiles, most students finished their learning process with memorizing and rehearsing. However, initial reading also occurred as a final activity in some mental learners (20%) and limited strategy users (9%), while this is not the case in the learning process of the other learning strategy profiles. In addition, some integrated strategy users (13%) finished their learning process with reflecting on their progress, while rereading also occurred as final activity in some limited strategy users (9%).

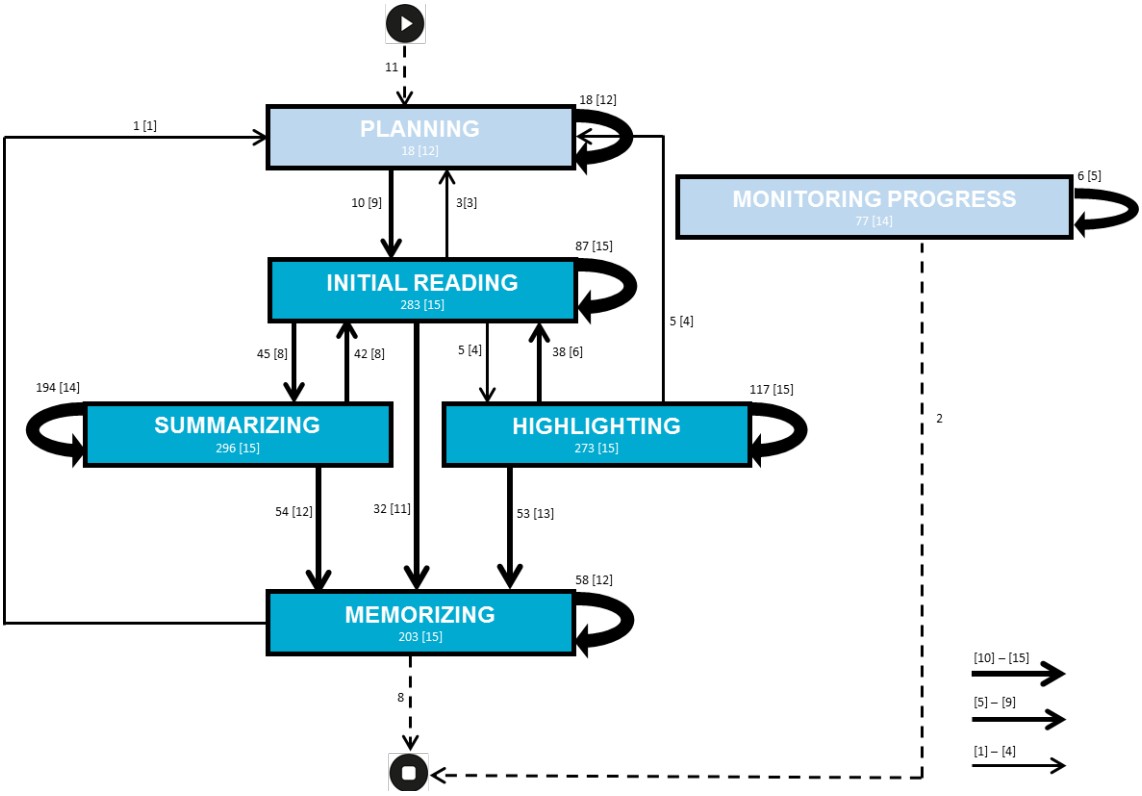


Figure 2

Text-learning process model of integrated strategy users (ISU; n = 15), including the frequencies of occurrence and, between brackets, the case frequencies (i.e., the number of students that conducted the activities at least once). The more frequent an activity was performed, the darker it is displayed. The more frequent a path between activities occurred, the thicker the arrow is displayed.

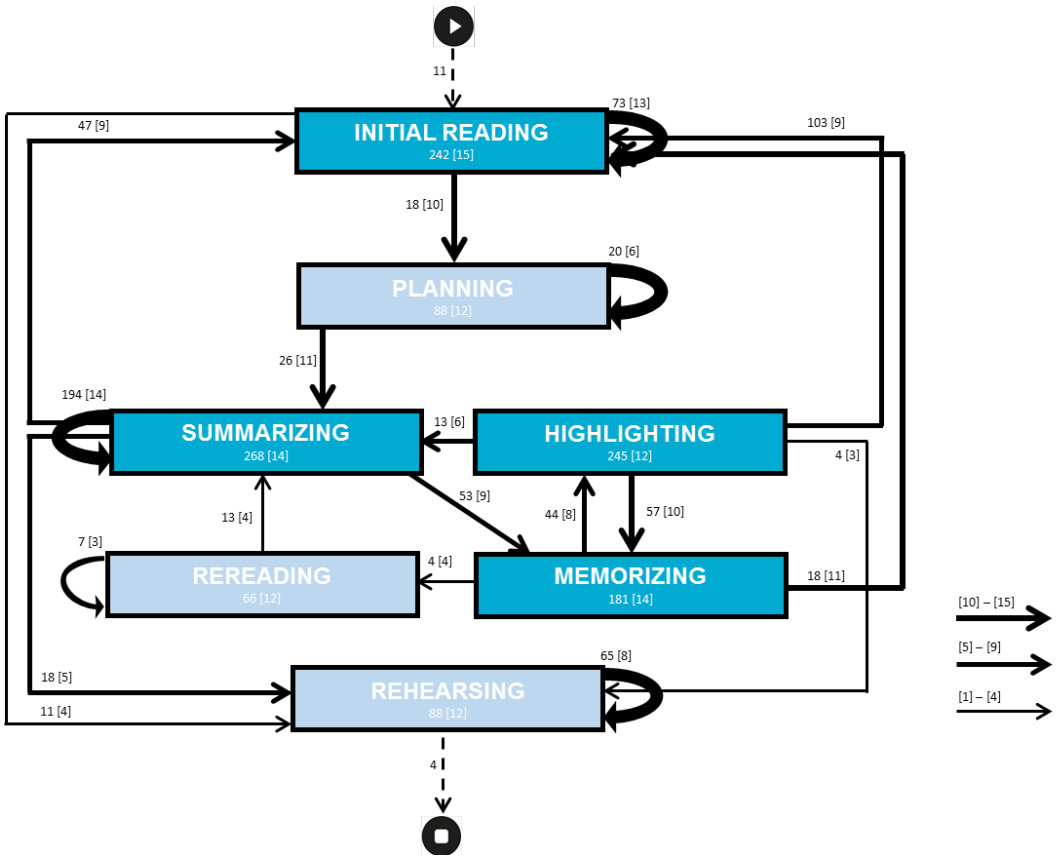


Figure 3

Text-learning process model of information organizers (IO; n = 15), including the frequencies of occurrence and, between brackets, the case frequencies (i.e., the number of students that conducted the activities at least once). The more frequent an activity was performed, the darker it is displayed. The more frequent a path between activities occurred, the thicker the arrow is displayed.

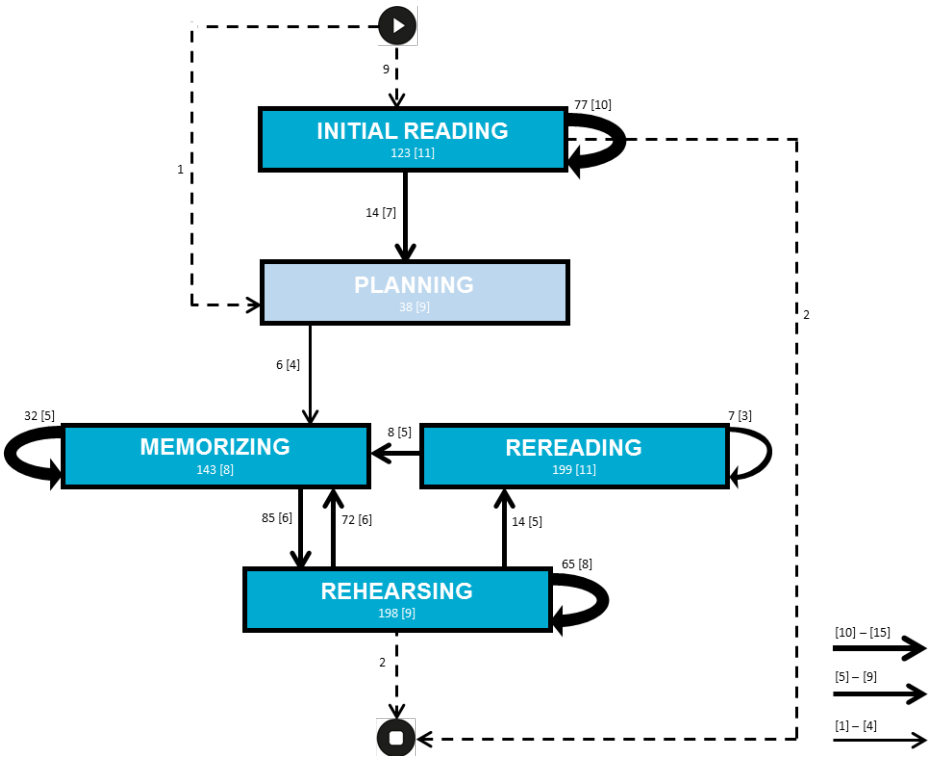


Figure 4

Text-learning process model of mental learners (ML; n = 10), including the frequencies of occurrence and, between brackets, the case frequencies (i.e., the number of students that conducted the activities at least once). The more frequent an activity was performed, the darker it is displayed. The more frequent a path between activities occurred, the thicker the arrow is displayed.

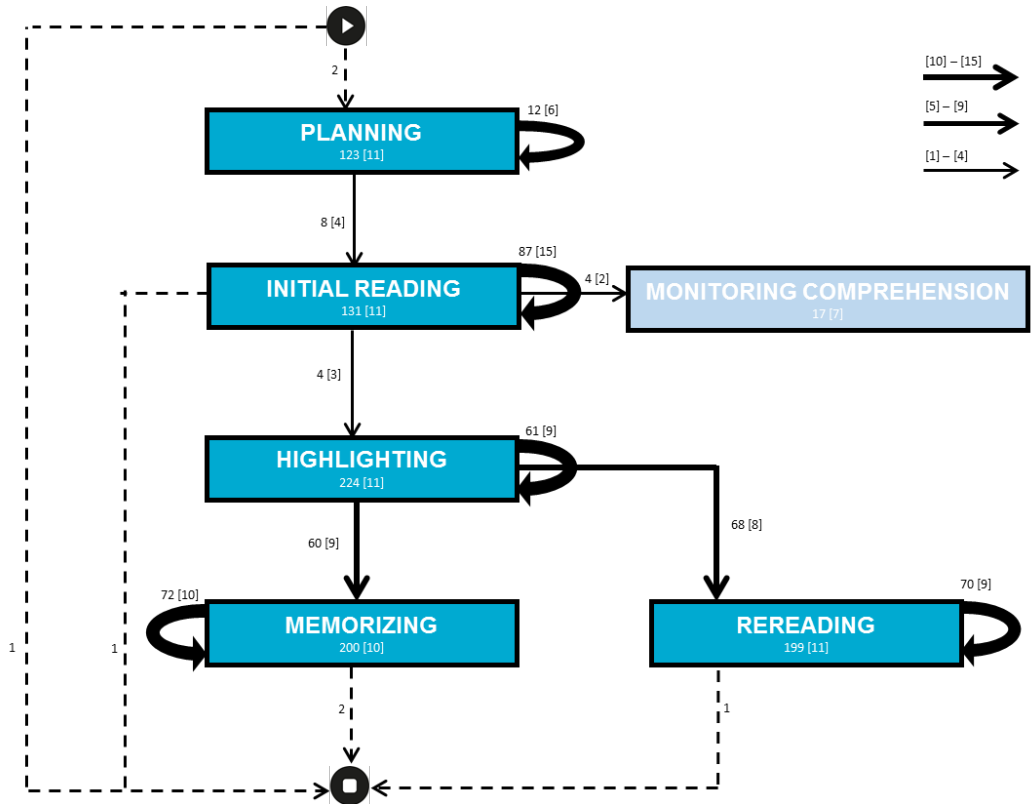


Figure 5

Text-learning process model of limited strategy users (LSU; n = 11), including the frequencies of occurrence and, between brackets, the case frequencies (i.e., the number of students that conducted the activities at least once). The more frequent an activity was performed, the darker it is displayed. The more frequent a path between activities occurred, the thicker the arrow is displayed.

4. Discussion

To date, little is known about the sequences in which certain strategies are applied by different learners during the course of their text-learning process. In this respect, it remains unclear whether a cyclical and flexible approach to learning, as put forward as the most effective in various theoretical models, unfolds in different learning strategy profiles when learning from text. Nevertheless, in-depth insight into students' learning processes enables to be responsive to individuals' learning needs and avoid 'one-size-fits-all' approaches to learning. The purpose of this study therefore was to uncover both the frequency of students' applied strategies throughout their learning process, as well as the temporal patterns between these text-learning strategies. More particularly, the strategic behaviour of students from four different learning strategy profiles (i.e., integrated strategy users, information organizers, mental learners, and limited strategy users) based on a retrospective self-report questionnaire in a previous study (Rogiers et al., 2019a), was further depicted and compared by means of educational process mining (EPM) on their think-aloud protocol (TAP) data. In this respect, both students' concurrently and retrospectively measured strategy use was complementary taken into account.

The first research question focused merely on the quantity of students' strategy use by studying the frequency wherein text-learning strategies were executed by the different learning strategy profiles during actual text learning. The results clearly correspond to the findings of Rogiers and colleagues (2019a) who determined different learning strategy profiles based on students' retrospectively self-reported text-learning strategies. The results postulated less diverse learning strategy use for limited strategy users (LSU; e.g., highlighting and rereading), versus more varied overt and covert text-learning strategies for integrated strategy users (ISU). Similarly, the frequent use of overt text-noting strategies (i.e., highlighting, summarizing) reported by information organizers (IO) was reflected in their verbalized learning behaviour. The same applies for mental learners (ML), who both reported and actually applied the frequent use of covert mental learning strategies (e.g., memorizing, rehearsing, paraphrasing). This was also reflected in the strategies included in the different process models (RQ2). In this regard, the clusters determined based on students' retrospective self-report data were largely confirmed by their concurrent TAP data. Although some research has clearly shown discrepancies between retrospective and concurrent measures of students' strategic behaviour, our comparison overall shows that both measures enable us to uncover which strategies students do or do not use frequently. Given this convergence, the current study provides empirical support for retrospective self-report questionnaires as acceptable alternatives for more time- and labour-intensive measures such as TAP (e.g., Greene & Azevedo, 2009). It must be noticed, however, that retrospective self-reports offer a more general picture of the frequency of students' strategy use, while TAP enable a more fine-grained analysis of students' learning process, for example by exploring the temporal patterns in which it unfolds. This was particularly elaborated on in response to the second research question by applying EPM.

With respect to the second research question (i.e., studying temporal patterns in students' text-learning process based on the sequences in which their strategies are applied), the process models of the learning strategy profiles enabled a qualitative and systematic analysis based on several theoretical models in the field (see introduction section). When overviewing the results regarding the second research question, three major aspects should be noticed.

First, the findings revealed that information organizers and mental learners immediately started their learning process with reading the text before considering their planning, whereas limited and integrated strategy users initiated their learning process with planning before they started to read. In addition, planning was strongly interwoven in the different phases of integrated strategy users' learning process (i.e., before, during and after learning), which was clearly different from the other process models. The connections with planning in integrated strategy users' process model could indicate that ISU adopted a more efficient and systematic study approach (Pintrich, 2000; Zimmerman, 2002). Pressley and colleagues (1987) for instance, consider good strategy users as planful strategy users who think before they act. Their plan is not conceived as a linear sequencing of strategies, however, but rather as interacting and integrating with other strategies throughout the learning process. While at a more basic level, learners will develop a single (reading) plan for reading text materials (for the reading task) before learning, more advanced learners additionally develop a profound (action) plan for task execution and learning (Desoete, 2007; Pressley, 2000), which was the case for integrated strategy users in the current study.

Second, a remarkable difference between the learning strategy profiles regards the use of monitoring strategies. On the one hand, progress monitoring was included in integrated strategy users' process model, suggesting that a considerable share of integrated strategy users actively tracked and controlled the quality of their progress and the available time left for task execution (Meijer et al., 2006; Moos & Azevedo, 2009). By applying this progress monitoring strategy, adherence to the plan is stimulated, as well as revisions to comply with the plan (Pressley et al., 1987). In this respect, progress monitoring during learning is strongly linked to planning before

and during learning. Hence, particularly in integrated strategy users' process model, metacognitive strategies (i.e., planning and progress monitoring) and, by extension, cognitive strategies, mutually interacted. On the other hand, comprehension monitoring was included in the process model of limited strategy users. This strategy refers to control activities directed at the correctness and comprehensiveness of one's understanding (Moos & Azevedo, 2009). An indicator of applying this strategy, for example, concerns learners' noting lack of full understanding, as well as efforts to monitor their understanding after reading the text (Veenman et al., 1997). As previous research shows that limited strategy users' level of reading ability is generally lower compared to the other learning strategy profiles, limited strategy users' level of reading ability may also have played a role here (Rogiers et al., 2019a).

Finally, the results showed that limited strategy users followed a rather linear sequenced approach, whereas integrated strategy users, mental learners, and particularly information organizers adopted a more cyclical approach to learning as they often repeated or returned to prior activities. Compared to the other learning strategy profiles, limited strategy users did not seem to interact with the text as actively and recursively. Instead, they confined their study behaviour to highlighting, memorizing, and rereading. Contrary, particularly information organizers and integrated strategy users frequently switched between various strategies throughout their learning process or resumed previous strategies. According to important theoretical models concerning successful strategy use (Alexander & Jetton, 2000; Pintrich, 2000; Pressley et al., 1987; Weinstein et al., 2011; Zimmerman, 2002), also the simultaneous use of different strategies is what characterises a good strategy user. As different strategies are executed ever more efficiently in good strategy users, Pressley and colleagues (1987) state that in these learners, more short-term capacity is 'left over' to adopt other strategies simultaneously and enhance their text learning (Pressley et al., 1987). In information organizers' process model, for instance, a large share of information organizers returned to reading the text or started memorizing or rehearsing the text after summarizing. Further, memorizing and highlighting were often performed in alternation and highlighting was also frequently followed by rehearsing, initial reading and/or summarizing. Remarkable, however, is that after conducting memorizing, summarizing, and highlighting activities, a large share of information organizers returned to initially reading the text. These paths could imply that information organizers started to engage in different text-learning strategies without first reading or fully understanding the study text. Equally, this might indicate that information organizers had the tendency to interrupt their first reading with other activities (Wade et al., 1990). This could be due to the fact that they did not initiate their actual learning process by planning this process in advance. In this respect, their text-learning process seems less systematic than, for example, integrated strategy users' learning process. Rather than directly selecting important ideas in the text or starting to summarize, the findings indicate that integrated strategy users read text fragments, deliberate on the importance of the given information and then highlighted or summarized the main ideas. Subsequently, integrated strategy users applied their notes as tools to memorize. This might again indicate that these learners adopted a more strategic approach to text learning. However, it is remarkable that integrated strategy users rarely applied rereading or rehearsing strategies during their text-learning process. Since they more actively monitored their progress, it might have been the case that they did not consider it necessary or feasible to repeat or rehears the text within the given time span. Further, also the recursive loop for memorizing, rehearsing, and rereading in mental learners' process model demonstrates an alternate application of these strategies. Although we must be aware of our small sample size, the initial reading activities as final activities in the process models of mental learners and limited strategy users could imply that some students finished their learning process quite abruptly and did not implement a thoughtful text-learning approach.

4.1 Limitations and implications

The present study is associated with some strengths and concerns regarding both the measure and data analysis approach used. First, we must be aware of the fact that various self-report measures reflect learning strategy conceptualizations in a different way. Retrospective self-report data has shown to be valuable in prior research to provide insight into the frequency and variety of applied overt, covert, cognitive, and metacognitive learning strategies during a learning task (e.g., Rogiers et al., 2019a; Merchie et al., 2014). However, this particular data provides us with less information on the cyclical and adaptive nature of these processes, characteristics that have been identified in various theoretical models as being essential in strategic learning. TAP can be regarded as concurrent self-report measures and are recognised as useful data sources to provide additional insight (Dinsmore, 2018; Veenman, 2011). More particularly, by the unique combination of concurrent self-report think-aloud data and educational process mining in this study, we were able to shed light on not only the diversity of applied learning strategies, both also on their cyclical and adaptive nature. In this way, EPM on students' concurrent TAP really opened the black box and provided in-depth insight into the course of different learners' actual text-learning process (Cromley & Wills, 2016; Veenman, 2005). In this respect, this study illustrates the complementarity of both retrospective (i.e., task-specific self-report questionnaires) and concurrent self-reports (i.e., TAP). This

reflection touches upon the first explicit question regarding self-report data tackled throughout the different contributions in this special issue. However, limitations of this study should equally be recognised. A first risk inherent to thinking out loud concerns the incompleteness as automated or unconscious behaviour is not explicitly verbalized (Boekaerts & Corno, 2005). It is possible that students' actions and thoughts might have sometimes remained covert, making them difficult to record in the TAP. Second, students were instructed to report both verbal and nonverbal processes during thinking out loud. To prevent that students' verbalisations did interfere with their learning process (Greene et al., 2011), they were not asked to explain these processes. Therefore, TAP gave no insight into students' underlying motives for their executed activities and sequences that occurred in the process models. Including retrospective stimulated interviews (Schellings & Broekkamp, 2011) based on students' TAP could be useful in future research to learn more about the underlying motives of students' behaviour.

As to the data analysis approach, no specific EPM guidelines are currently available with respect to the number of activities and paths to be included in the process models. More process mining research in educational settings, as well as exploring EPM techniques that rely on different algorithms could therefore contribute to a better understanding of students' learning process on the one hand and to more evidence-based guidelines for conducting EPM on the other hand (Bannert et al., 2014). Related to this, it is to be recommended as well to engage in more fine-grained coding of the think aloud data in future research in view of considering valences of specific SRL processes during students' learning. Positive judgment of learning (e.g., "I am getting this"), for example, can elicit a distinct subsequent SRL process than a negative judgment of learning (e.g., "I am so confused with this paragraph"). Within the scope of the current study, however, this fine-grained coding was not applied (e.g., both 'detecting lack of comprehension' and 'mentioning awareness of understanding' were more generally coded as 'monitoring comprehension'). We therefore make a plea for more fine-grained coding of the distinct subprocesses of particular learning strategies, such as for instance 'comprehension monitoring', to enable the study of more detailed subprocesses and their temporal nature.

Further, prior studies have shown that students adapt their strategies and the effort they spend on studying according to the learning task, their prior domain knowledge, and their learning goals (e.g., Boekaerts & Niemivirta, 2000; Broekkamp & Van Hout-Wolters, 2007). As a result, students may decide to select from their available strategies these strategies that are most appropriate given the assigned learning task, their prior knowledge, and/or the learning goal they set for themselves. In this respect, they might opt, for example, to systematically reread the study text instead of engaging in summarizing and paraphrasing the text (Broekkamp & Van Hout-Wolters, 2007). It will therefore be interesting in future research to study students' strategy use across more and varied learning tasks as well as to investigate the impact of their prior knowledge and their personal learning goals. In this respect, it is to be recommended to also consider other types of prior knowledge tests, such as open questions, multiple choice tests, cloze tests, completion tests, and recognition tests, which also provide valid means of assessment (Dochy et al., 1999).

With regard to the implications for research, this study extends earlier work by including new possibilities for analysing learning processes by means of EPM. This study must be considered as a first important investigation in unravelling patterns in secondary school students' text-learning processes. Educational research is encouraged to fine-tune this type of analysis by the suggestions mentioned above. Further, our results encourage data triangulation in future research, preferably combining both off-line (e.g., self-report questionnaires) and on-line measures (e.g., TAP) to gain a more accurate portrayal of students' learning process (Rogiers et al., 2019b; Veenman, 2005). In view of implications for theory, this study showed that EPM can be used to test the cyclical and adaptive use of learning strategies as put forward in different theoretical models. The present study confirmed differences between four learning strategy profiles in secondary school students. These findings carry important implications for educational practice to help and support students to also evolve towards the more adaptive and cyclical use of strategies. The proposed process models provide a detailed picture on students' text-learning process and could be used as starting points for supporting learner-oriented teaching and learning.

Keypoints

- Integrated strategy users execute more diverse text-learning strategies than the other learner profiles
- Limited and integrated strategy users initiate their learning process with planning this process, while planning is strongly interwoven in the different phases of integrated strategy users' learning process
- Progress monitoring is included in integrated strategy users' process model, comprehension monitoring is included in limited strategy users' process model
- Limited strategy users follow a rather linear sequenced approach to learning whereas integrated strategy users, mental learners, and particularly information organizers adopt a more cyclical approach to learning

References

Alexander, P. A. (1998). The nature of disciplinary and domain learning: The knowledge, interest, and strategic dimensions of learning from subject matter text. In C. R. Hynd (Ed.), *Learning from texts across conceptual domains* (pp. 263-287). New York, NY: Routledge.

Alexander, P. A., & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 285-310). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Alexander, P. A., Grossnickle, E. M., Dumas, D., & Hattan, C. (2018). A retrospective and prospective examination of cognitive strategies and academic development: Where have we come in twenty-five years? In A. O'Donnell (Ed.), *Oxford Handbook of Educational Psychology.* Oxford, UK: Oxford University Press.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning, 9*(2), 161–185. https://doi.org/10.1007/s11409-013-9107-6

Bergman, L. (2001). A person approach to adolescence: Some methodological challenges. *Journal of Adolescent Research, 16*(1), 28–53. https://doi.org/10.1177/0743558401161004

Boekaerts, M., & Corno, L. (2005). Self regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199–231. https://doi.org/10.1111/j.1464-0597.2005.00205.x

Boekaerts, M., & Niemivirta, M. (2000). Self-regulated learning: Finding a balance between learning goals and ego-protective goals. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation*. San Diego, CA: Academic Press.

Bråten, I., & Samuelstuen, M. S. (2004). Does the influence of reading purpose on reports of strategic text processing depend on students' topic knowledge? *Journal of Educational Psychology, 96*(2), 324-336. https://doi.org/10.1037/0022-0663.96.2.324

Båten, I., & Samuelstuen, M. S. (2007). Measuring strategic processing: comparing task-specific self-reports to traces. *Metacognition and Learning, 2*(1), 1-20. https://doi.org/10.1007/s11409-007-9004-y

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cromley, J. G., & Wills, T. W. (2016). Flexible strategy use by students who learn much versus little from text: transitions within think-aloud protocols. *Journal of Research in Reading, 39*(1), 50-71. https://doi.org/10.1111/1467-9817.12026

Deed, C., Lesko, T. M., & Lovejoy, V. (2014). Teacher adaptation to personalized learning spaces. *Teacher Development, 18*(3), 369–383. doi:10.1080/13664530.2014.919345

Dinsmore, D. L. (2018). *Strategic Processing in Education*. New York, NY: Routledge.

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*(2), 145-186. https://doi.org/10.3102/00346543069002145

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215- 251. https://doi.org/10.1037//0033-295x.87.3.215

Fluxicon. (2019). *Disco user's guide*. Retrieved from https://fluxicon.com/disco/files/Disco-UserGuide.pdf

Fox, E. (2009). The role of reader characteristics in processing and learning from informational text. *Review of Educational Research, 79*(1), 197-261. https://doi.org/10.3102/0034654308324654

Fryer, L. K., & Vermunt, J. D. (2017). Regulating approaches to learning: Testing learning strategy convergences across a year at university. *British Journal of Educational Psychology*, *88*(1), 21-41. https://doi.org/10.1111/bjep.12169

Günther, C., & Van der Aalst, W. (2007). Fuzzy mining: Adaptive process simplification based on multi-perspective metrics. In G. Alonso, P. Dadam, & M. Rosemann (Eds.), *International conference on business process management (BPM 2007)* (pp. 328–343). Berlin, Germany: Springer.

Greene, J. A., Robertson, J., & Croker Costa, L.-J. (2011). Assessing self-regulated learning using think-aloud methods. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 313–328). New York, NY: Routledge.

McNamara, D. S., Ozuru, Y., Best, R., & O'Reilly, T. (2007). The 4-pronged comprehension strategy framework. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theories, Interventions, and Technologies* (pp. 465-496): New York, NY: Erlbaum

Merchie, E., & Van Keer, H. (2014). Learning from Text in Late Elementary Education. Comparing Think-aloud Protocols with Self-reports. *Procedia - Social and Behavioral Sciences, 112*(2013), 489–496. https://doi.org/10.1016/j.sbspro.2014.01.1193

Merchie, E., Van Keer, H., & Vandevelde, S. (2014). Development of the Text-Learning Strategies Inventory: Assessing and profiling learning from texts in fifth and sixth grade. *Journal of Psychoeducational Assessment, 32*(6), 1-15. https://doi.org/10.1177/0734282914525155

Organisation for Economic Cooperation, Development [OECD]. (2006). Personalizing Education. Paris: OECD Publishing.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*(422), 1–28. https://doi.org/10.3389/fpsyg.2017.00422

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

Pressley, M., Borkowski, J. G., & Schneider, W. (1987). Cognitive strategies: Good strategy users coordinate metacognition and knowledge. *Annals of Child Development*, *4*, 89-129.

Reimann, P. (2007). Time is precious: Why process analysis is essential for CSCL (and can also help to bridge between experimental and descriptive methods). In C. Chinn, G. Erkens & S. Puntambekar (Eds.), *Mice, minds, and society. Proceedings of the computer-supported collaborative learning conference (CSCL 2007)* (pp. 598–607). New Brunswick, NJ: International Society of the Learning Sciences.

Rogiers, A., Merchie, E., & Van Keer, H. (2018). *Fostering text-learning strategies in secondary education through explicit strategy-instruction.* Paper presented at the International Conference of the EARLI SIG 2, Freiburg, Germany, 27-29 august, 2018.

Rogiers, A., Merchie, E., & Van Keer, H. (2019a). Learner profiles in secondary education: Occurrence and relationship with performance and student characteristics. *The Journal of Educational Research, 112*(3), 385-396. https://doi.org/10.1080/00220671.2018.1538093

Rogiers, A., Merchie, E., & Van Keer, H. (2019b). What they say is what they do? Comparing task-specific self-reports, think-aloud protocols, and study traces for measuring secondary school students' text-learning strategies. *European Journal of Psychology of Education*, 1-18. https://doi.org/10.1007/s10212-019-00429-5

Samuelstuen, M. S., & Bråten, I. (2007). Examining the validity of self-reports on scales measuring students' strategic processing. *British Journal of Educational Psychology, 77*(2), 351-378. https://doi.org/10.1348/000709906X106147

Schellings, G. L. M., & Broekkamp, H. (2011). Signaling task awareness in think-aloud protocols from students selecting relevant information from text. *Metacognition and Learning, 6,* 65–82. https://doi.org/10.1007/s11409-010-9067-z

Slotte, V., Lonka, K., & Lindblom-Ylanne, S. (2001). Study-strategy use in learning from text. Does gender make any difference? *Instructional Science, 29*(3), 255-272. https://doi.org/10.1023/a:1017574300304

Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Computers in Human Behavior, 28*(4), 1321–1331. https://doi.org/10.1016/j.chb.2012.02.016

Scott, D. B. (2008). Assessing text processing: A comparison of four methods. *Journal of Literacy Research, 40*, 290-316. doi:10.1080/10862960802502162

Simpson, M. L., & Nist, S. L. (2000). An update on strategic learning: It's more than textbook reading strategies. *Journal of Adolescent & Adult Literacy*, *43*(6), 528-541.

Stoeger, H., Fleischmann, S., Obergriesser, S. (2015). Self-regulated learning (SRL) and the gifted learner in primary school: The theoretical basis and empirical findings on a research program dedicated to ensuring that all students learn to regulate their own learning. *Asia Pacific Education Review,* 16, 257-267. https://doi.org/ 10.1007/s12564-015-9376-7

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artett & B. Moschner (Eds.), *Ledrnstrategien und metakognition. Implikationen für forschung und praxis* (pp. 77-99). Münster: Waxmann.

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: A discussion. *Metacognition and Learning, 6*(2), 205–211. https://doi.org/10.1007/s11409-011-9080-x

Veenman, M. V., Elshout, J. J., & Meijer, J. (1997). The generality vs. domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction, 7*(2), 187–209. https://doi.org/10.1016/S0959-4752(96)00025-4

Wade, S. E., Trathen, W., & Schraw, G. (1990). An analysis of spontaneous study strategies. *Reading Research Quarterly, 25*(2), 147-166. https://doi.org/10.2307/747599

Weinstein, C. E., Jung, J., & Acee, T.W. (2011). Learning strategies In V. G. Aukrust (Ed.), *Learning and Cognition in Education* (pp. 137-143). Oxford, UK: Elsevier Limited.

Van der Aalst. (2011). *Process mining: Discovery, conformance and enhancement of business processes.* New York, NY: Springer.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method. A practical guide to modelling cognitive processes.* London, UK: Academic Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*(2), 64–71.

Young, K. A. (2005). Direct from the source: the value of 'think-aloud' data in understanding learning. *Journal of Educational Enquiry, 6*(1), 19-33.

Appendices

Appendix A

Descriptive statistics and reliability coefficients of the different Text-Learning Strategies Inventory subscales

TLSI-subscales	***N* items**	**Example item**	***M (SD)***	**Cronbach's α**
Summarizing and schematizing (SS)	7	*I repeated the text with my summary or graphic organizer on my scratch paper*	3.14 (1.30)	.92
Highlighting (HL)	1	*I marked the most important things*	4.31 (1.31)	/
Rereading (RR)	3	*To learn the text, I read the text a lot of times*	3.14 (1.07)	.72
Paraphrasing (PAR)	7	*I covered up the text information and tried to recall it*	3.02 (0.82)	.71
Linking with prior knowledge (LPK)	3	*Before learning, I thought about what I already knew*	3.16 (1.09)	.75
Studying titles and pictures (TP)	3	*I looked at the titles to understand the text*	2.81 (1.07)	.70
Planful approach (PA)	3	*First, I read the whole text and then I started learning*	3.86 (1.08)	.65
Self-evaluation (SE)	5	*While learning, I checked what I had already done and how much I still had to do*	4.03 (0.63)	.70
Monitoring (MON)	5	*I managed to learn the text in a good way*	3.21 (0.88)	.63

Note. TLSI = Text-Learning Strategies Inventory. Cronbach's α is based on the total sample of 1,931 students wherein learner profiles were determined.

Appendix B

Results of the Post Hoc Pairwise comparisons between the four learning profiles on the different coding categories

Coding category		**IO**		**ML**		**LSU**		
Overt cognitive strategies		***p***	**Cohen's *d***	***p***	**Cohen's *d***	***p***	**Cohen's *d***	***F***
Summarizing	ISU	.128	.075	.000***	.700	.000***	.403	$F(3, 4732) = 85.59$, $p < .001$
	IO			.000***	.631	.000***	.329	
	ML					.000***	-.351	
Highlighting	ISU	.291	.077	.000***	.511	.951	.000	$F(3, 4732) = 43.88$, $p < .001$
	IO			.000***	.435	.288	-.076	
	ML					.000***	-.510	
Covert cognitive strategies		***p***	**Cohen's *d***	***p***	**Cohen's *d***	***p***	**Cohen's *d***	***F***
Memorizing	ISU	.720	.058	.987	-.028	.183	-.081	$F(3, 4732) = 4.54$, $p = .004$
	IO			.373	-.086	.002**	-.139	
	ML					.893	-.053	
Rereading	ISU	.001***	-.164	.000***	-.208	.006**	.199	$F(3, 4732) = 22.89$, $p < .001$
	IO			.996	-.046	.000***	.335	
	ML					.000***	.372	
Rehearsing	ISU	.013*	-.141	.000***	-.617	.835	.063	$F(3, 4732) = 144.34$, $p < .001$
	IO			.000***	-.494	.000***	.205	
	ML					.000***	.671	
Paraphrasing	ISU	.023*	.160	.000***	-.193	.045*	.083	$F(3, 4732) = 21.12$, $p < .001$
	IO			.000***	-.356	.000***	-.233	

	ML					.156	.107	
Elaborating	ISU	.004**	.184	.978	.064	.991		*F*(3, 4732) = 5.252, *p* = .001
	IO			.226	-.080	.004**	-.146	
	ML					.874	-.063	
Metacognitive strategies		***p***	**Cohen's *d***	***p***	**Cohen's *d***	***p***	**Cohen's *d***	***F***
Planning	ISU	.000***	-.264	.010*	-.180	.000***	-.423	*F*(3, 4732) = 38.56, *p* < .001
	IO			.438	.089	.000***	-.178	
	ML					.000***	-.264	
Monitoring progress	ISU	.000***	.210	.000***	.280	.232	.093	*F*(3, 4732) = 15.57, *p* < .001
	IO			.995	.082	.016*	-.116	
	ML					.001***	-.189	

Note. *$p \leq .05$, ** $p \leq .01$, ***$p \leq .001$. To interpret effect sizes, Cohen's benchmarks for the social sciences apply (i.e., small effect size: $d = 0.2$, medium effect size: $d = 0.5$, and large effect size: $d = 0.8$; Cohen, 1977). *ISU* = integrated strategy users, *IO* = information organizers, *ML* = mental learners, *LSU* = limited strategy users.

Rossi, P., Wright, J., & Anderson, A. (1983). *Handbook of survey research. Sample surveys: History, current practice, and future prospects*. San Diego: Academic Press.

Schellings, G. (2011). Applying learning strategy questionnaires: problems and possibilities. *Metacognition and Learning, 6*(2), 91-109. https://doi.org/10.1007/s11409-011-9069-5

Schellings, G., & Van Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments: theoretical and empirical considerations. *Metacognition and Learning, 6*(2), 83-90. https://doi.org/10.1007/s11409-011-9081-9

Schwarz, N. (1990). Assessing frequency reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. In *Research methods in personality and social psychology.* (pp. 98-119). Thousand Oaks, CA, US: Sage Publications, Inc.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*(2), 277-287. https://doi.org/10.1002/acp.1340

Singleton, R., & Straits, B. (2009). *Approaches to social research - 5th edition*. Oxford: Oxford University Press.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics*: Oxford University Press.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Unsworth, N., Heitz, R., Schrock, J., & Engle, R. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505. https://doi.org/10.3758/bf03192720

van Gog, T., & Jarodzka, H. (2013). Eye Tracking as a Tool to Study and Enhance Cognitive and Metacognitive Processes in Computer-Based Learning Environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143-156). New York: Springer.

Veenman, M. (2011). Alternative assessment of strategy use with self-report instruments: a discussion. *Metacognition and Learning, 6*(2), 205-211. https://doi.org/10.1007/s11409-011-9080-x

Veenman, M., & van Hout-Wolters, B. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und Metakognition: Implikationen für Forschung und Praxis* (pp. 77-99): Münster: Waxmann.

Vermunt, J., & Donche, V. (2017). A learning patterns perspective on student learning in higher education: state of the art and moving forward. *Educational psychology review, 29*(2), 269-299. https://doi.org/10.1007/s10648-017-9414-6

Willis, G., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods, 23*(4), 331-341. https://doi.org/10.1177/1525822X11416092

Willis, G., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology, 5*(3), 251-267. https://doi.org/10.1002/acp.2350050307

Yeari, M., Oudega, M., & van den Broek, P. (2016). The effect of highlighting on processing and memory of central and peripheral text information: evidence from eye movements. *Journal of Research in Reading, 40*(4), 365-383. https://doi.org/10.1111/1467-9817.12072

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 63 - 84
ISSN 2295-3159

Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design

Julia Moeller[a], Jaana Viljaranta[b], Bärbel Kracke[c], & Julia Dietrich[c]

[a] University of Leipzig, Germany,
[b] University of Eastern Finland, Joensuu, Finland,
[c] University of Jena, Germany

Article received 25 June / revised 21 December / accepted 21 October / available online 30 March

Abstract

This article proposes a study design developed to disentangle the objective characteristics of a learning situation from individuals' subjective perceptions of that situation. The term objective characteristics refers to the agreement across students, whereas subjective perceptions refers to inter-individual heterogeneity. We describe a novel strategy for assessing and disentangling objective situation characteristics and subjective perceptions thereof, propose methods for analysing the resulting data, and illustrate the procedure with an example of a first study using this design to examine situational interest in 155 university students. Situational interest was assessed nine times per weekly lecture with three measurement time points per person and a rotated multi-group schedule. Assessments took place over the course of an entire semester of ten weeks.

One of the advantages of the proposed design is that objective group agreements can be disentangled from subjective deviations from the group's average at each of the nine measurement time points per weekly lecture. Furthermore, the proposed design makes it possible to study the development of both subjective and objective parameters across the time span of one weekly lecture and an entire semester, while the burden for each person is kept relatively low with three beeps per lecture.

Keywords: subjective self-reports, inter-rater agreement, experience sampling method, momentary motivation.

Corresponding author: Julia Moeller, Marschnerstr. 31, 04109 Leipzig, Germany, julia.moeller@uni-leipzig.de DOI: https://doi.org/10.14786/flr.v8i3.529

1. Introduction

Imagine you attend a lecture that you love but that all your classmates seem to hate, dread, or find boring. You just love this lecture of Statistics and Research Methods, because of its exciting implications for epistemology and its answers to the question where knowledge comes from and how much we can(not) trust in what we know. You think this is one of the best, most interesting courses you have ever had, but your fellow students just don't seem to share your enthusiasm for the philosophy of science or mathematical representation of knowledge. While you express your love for this Statistics and Methods course, nearly everyone else would rather study "real Psychology" or sleep in instead of starting the day with the 8:00 a.m. Statistics course. One of your classmates even called you a nerd. While you try to convince everyone that this course is *objectively interesting*, the other students try to convince you that this is an *objectively uninteresting* lecture, claiming that "if we all agree it's boring, it can't be *objectively interesting*". Your best friend agrees with the others on that, but, trying to put himself in your shoes, also acknowledges that you have *subjective reasons* to find that lecture interesting, while also trying to convey to you that your *subjective interest* just isn't everyone's cup of tea. You discuss to what extent the agreement, or average interest, of the class reflects the *objective* interestingness of that course.

The Dean, in turn, holds the teaching evaluation in hand when announcing the decision to discontinue your favourite course in the future, citing the average lack of interest of the attendants as evidence for the *objective lack of teaching quality*, because all other Psychology courses got higher student ratings in the questions asking about students' enthusiasm. You feel unheard and unseen, after all, aren't you a data point in that statistic the Dean holds in hand, too? Didn't your favourite Statistics teacher just yesterday teach you about the problem that sometimes individual students or subgroups hide behind the overall trend, so that we need methods to detect and describe these subgroups and deviating individuals?

This article presents a novel approach to disentangle and describe both the overall trend in the agreement of a class on the ratings of a learning situation, and the deviations of individual, subjective, perceptions from that overall trend. The methods proposed in this article promise to be insightful for a broad audience, including researchers using the experience sampling method for classroom assessments, educational technology developers looking for methods to provide metrics and visuals concerning student heterogeneity and objective situation characteristics in teacher dashboards and class feedback systems, as well as educators who are interested in situational measures of students' classroom perceptions, momentary assessments supporting personalised learning, or teacher feedback for social-emotional learning. While we use the example of interest ratings throughout the article, the methods proposed here could also be applied to assess other learning-related classroom perceptions, such as students' observations of teacher behaviour, or students' perceptions of the current task being difficult or easy, to name a few. By disentangling the idiosyncratic and commonly shared components of motivational self-reports, this article makes a contribution to this special issue's first question ("In what ways do self-report instruments reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use?"). We also address the second question of this special issue by proposing analytics strategies, but rather than focusing on the constraints mentioned in the special issue editorial, we focus on novel avenues for analyses.

1.1 How to assess characteristics of learning situations

Situational self-report assessments of motivation and emotion are more and more frequently used, thanks to new technology that makes it easier and cheaper than ever to ask participants in real-time via mobile devices about their current activities, as well as their subjective perceptions, feelings, and motivations, pertaining to these currently ongoing activities. The methods used to gather such data are called Experience Sampling Method (ESM; e.g., Hektner et al., 2007), ambulatory assessments (e.g., Fahrenberg, 1996), or ecologically momentary assessments (Shiffman et al., 2008). In this article, we use the term ESM.

Compared to the classic retrospective one-time administered self-report questions for motivation and emotions, ESM assessments have several advantages: A first advantage is that ESM

assessments can capture the fluctuating and situation-specific components of motivation and emotions, while common retrospective, one-time administered self-reports do not reveal which aspects of the assessed variables fluctuate or remain stable from one situation to another. It is even possible to disentangle situational determinants (e.g., the exciting learning video used in today's lecture) from stable personal factors (e.g., this student's well-developed personal interest in the topic taught today or this students' general openness to experience), or contextual factors (e.g., the generally monotonous teaching style of this teacher, or the loud noise in that classroom from the construction side next door, which has hampered the students' attention and motivation for a year now). To disentangle such situational, personal, and contextual influences, ESM assessments can be combined with multilevel data analysis that decomposes the variance on the situational (within) level from the variance due to stable inter-individual differences (between level 1) and the variance between in contexts, such as class or school (between level 2; see e.g., Dietrich et al., 2017; Ketonen et al., 2018).

A second advantage is that situational measures have been discussed to be more valid than the retrospective self-reports, because in-the-moment assessments can reduce memory errors (e.g., Green et al., 2006; Takarangi et al., 2006) and response biases linked to beliefs and stereotypes that are otherwise activated in certain retrospective self-reports (e.g., Bieg et al., 2015; Goetz et al., 2013). While retrospective measures require participants to mentally aggregate their *typical* experience across all the situations they can remember, ESM data enable the researcher to empirically calculate such an aggregated typical experience as the mean score of the many repeated situational assessments for each person.

These advantages of ESM measures notwithstanding, they are still self-reports and therefore share many of the shortcomings related to self-reports with the retrospective measures. One of these shortcomings is the problem that ESM and other self-report measures capture only the subjective perception and rating of an experience, which do not necessarily reflect how other students would perceive the same situation. For example, if a student indicates a current interest of "4 – *very much*" on a four-point scale, the reasons for that choice of this response option remain unclear. Did this student choose this rating because he/she had a personal interest in this topic, while most other students were utterly bored? Or because this was the most captivating topic ever taught, and every student in the class was captivated and would agree? Or did the student just affirm being interested because of a very high individual level of trying to appear socially desirable? In classic ESM studies, it is very difficult to disentangle these different options, because typically, students are asked about activities at random times, implying that every student has an own individual random survey schedule, so that there is usually no way to determine how other students perceived the exact same situation.

To provide a solution for that problem, this study presents a research design that enables researchers to systematically assess groups of students at the same time points, so that inter-individual agreements and subjective deviations from these agreements can be distinguished.

The previous literature provides some examples of theories that distinguish between the subjective and objective components of information provided by self-reports in education (e.g., Göllner Wagner et al., 2018; Lüdtke et al., 2009). One such example is the research on interest, particularly the person-object theory of interest (Fink, 1991; Krapp, 2002; Krapp & Fink, 1992; Prenzel et al., 1986), which distinguishes between objective characteristics of a learning situation and the individual's subjective perceptions thereof, as well as distinguishing between fluctuating situational and stable personal determinants of the subjective perceptions. As the name person-object theory suggests, interest is expected to result from two main conditional factors, person characteristics and situation characteristics (Krapp, 1998). According to this theory, interest emerges in the interaction of a person with particular objects (including concrete objects, such as texts, and more abstract ideas, events, topics, texts, etc.). Objects are expected to differ in their likelihood of eliciting situational interest in individuals, depending on their verifiable, observable features. For instance, learning materials are more likely to trigger situational interest if their objective features make them surprising, novel, visually stimulating, and intense for the students. Texts are likely to trigger situational interest if they are, for instance, easy to comprehend, cohesive, vivid, if they evoke emotional reactions, and allow for collaborations with others (for overviews, see e.g., Krapp et al., 1992).

The (objective) interestingness of a situation is then perceived by individuals who differ in their stable, dispositional personal interests (Krapp et al., 1992) and their perceptions of the situation (e.g., a given information about the theory of relativity might be new to most students in a class, except for Max, who has heard about the topic extensively over dinner from his mother, who is a Quantum Physics professor). The individuals then feel more or less interested in the current learning situation, depending on their previous dispositional interest in the currently discussed topic (person characteristic) and the learning situations' objective characteristics. Thus, a fluctuating psychological state of more or less situational interest can be observed, which is either an expression of the currently actualised dispositional interest, or the fluctuating reaction to the objectively interesting situation, or a mixture of both (e.g., Krapp, 1998). The distinction between objective situation characteristics, student's subjective perceptions of these objective situation characteristics, and objective person characteristics is depicted in Figure 1, based on the person-object theory of interest visualized in Krapp (1998).

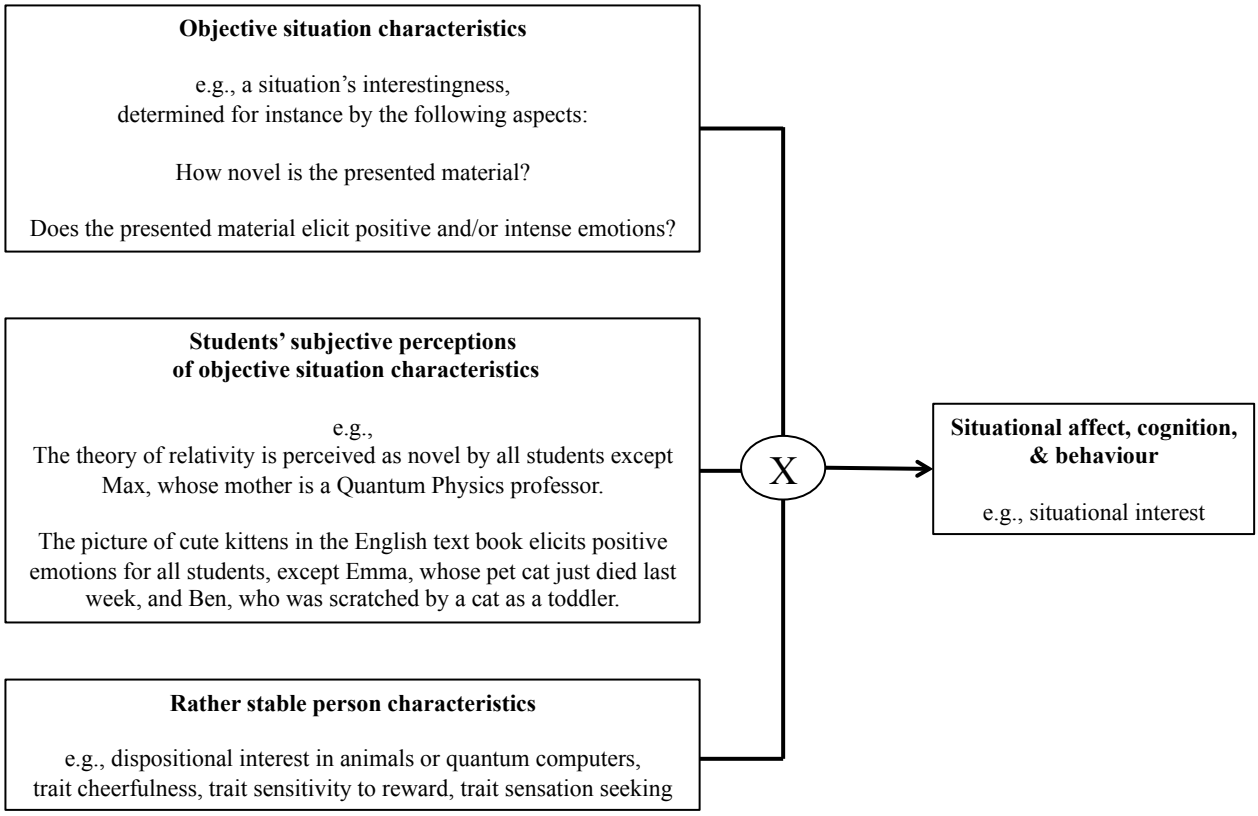


Figure 1. A model of the person-object theory of interest based on Krapp, (1998). We added the distinction between objective situation characteristics and subjective perceptions thereof.

We expect that such models of person-object interactions will be fruitful not only for the understanding of interest, but also for the understanding of other motivational constructs, cognitions, and behaviours. For example, a student's overall perception of a teacher's behaviour in a given learning situation may be influenced by objective situation characteristics (e.g., the teacher's in-fact behaviour), and by the student's subjective perceptions of the teacher's objective behaviour, and by stable or fluctuating student characteristics (e.g., the student's subjective liking of that teacher or the student's dispositional interest in the subject). Likewise, a student's learning in a given learning situation may be influenced by objective situation characteristics (e.g., the difficulty of the task, compared to tasks previously presented to the same student), by the student's subjective perceptions of those objective situation characteristics (e.g., the student's subjective appraisals, self-efficacy), as well as by stable person characteristics (e.g., the student's intelligence, ability self-concept, perseverance in the face of obstacles). These are just a few of the possible applications of the person-object-logic presented in

Figure 1, suggesting that this model might be a useful framework for understanding and assessing learning processes in classrooms.

Please note that we use the term "stable" in the following to refer to aspects that do not change across measurement time points for the duration of an experience sampling method study (typically a few days), meaning aspects that are modelled in multilevel models on the person-level as stable inter-individual difference components. By using the term "stable" in that sense, we do not mean to imply that those empirically stable components cannot develop over longer periods of time, and we do not mean to rule out that personality development takes place. We merely imply that such eventual long-term development is usually not captured or distinguishable by the short-term longitudinal data that we discuss in this article.

1.2 The present research

This article has three main goals: First, we introduce a novel design for ESM studies that enables researchers to disentangle the objective characteristics of a situation (e.g., the situation's interestingness, in terms of the inter-subjective agreement of all students) from each person's subjective perception of that situation (e.g., the subjective interest and deviation from the aforementioned inter-subjective agreement). Second, we propose different analytical strategies to analyse data assessed with this novel design and illustrate some of these proposed analyses using data from a first study that has used this novel assessment design to assess situational motivation in 155 university students. Third, we discuss a number of possible additional strategies of contrasting the situational self-reports of ESM assessments with more objective data, such as behavioural classroom observations or psychophysiological measures of emotion-related data.

This is a theoretical article with the main goal to propose a new assessment design, and to discuss its advantages as well as limitations. Therefore, our research questions refer to theoretical and methodological rationales, while the empirical results presented in this article merely serve as illustration and example for the methodological discussion, rather than being a centrepiece. Thus, the topic examined in the empirical part (students' situational interest in higher education) is treated as a rather exchangeable example of a construct, which to examine the here proposed method could help.

1.3 Research questions

RQ1: What concepts of objectivity can be applied to disentangle objective characteristics of situations from participants' subjective perceptions thereof?

RQ2: How do ESM research designs and schedules have to look like in order to capture both the objective situation characteristics and the subjective perceptions thereof?

RQ3: What analyses are needed to disentangle the objective situation characteristics and the subjective perceptions thereof in data collected with a design proposed under RQ2?

The research questions are mostly answered theoretically, but with references to an empirical study that will serve as an illustrating example for the proposed methods. This example study is described in the following.

2. Methods

2.1 Sample and procedure

The participants were 155 German university students (51% female; mean age $M = 21.77$ years, $SD = 2.91$; range: 19 to 46 years). The participants studied in a teacher education program with the aim to become subject teachers for secondary schools. Students provided intensive longitudinal data in the form of ESM surveys and were followed over one semester in a weekly lecture with 90-minute lessons

(except for lesson 4, which ended after 60 minutes ahead of schedule). The subject of the course was 'Psychological fundamentals of learning'. In each of ten consecutive weeks, students received notifications and questionnaires at fixed schedules, three times during each lesson, consisting of ten situational motivation items. The participants chose whether to respond online with their own smartphone or on paper-and-pencil questionnaires (smartphone: 58–71% participants with a mean of 65% across the ten lessons; paper-and-pencil: 29–42% participants with a mean of 35%).

N = 155 students provided valid information on situational measures in at least one lesson. During the data cleaning, we removed responses in the following cases: if the response was given more than 15 minutes after the signal (applies to the time-stamped online responses, not the paper-and-pencil response); if a person reported being present at the lecture but responded online after the lecture had ended; if a person responded to the three surveys immediately after another; and if a person responded with the same value on all ten items. This resulted in the omission of 251 surveys. A total of 2,226 completed ESM surveys remained in the analysis sample, which equals 48.94% of the possible full data (three responses per lesson by ten lessons, except for week four, which ended early and therefore included two responses per person only) by 155 participants resulting in 4,495 responses). 2200 of those completed ESM surveys had valid responses on at least one of the interest variables used in the analysis for this article and thus appear as the sample size in our Mplus output (Moeller et al., 2019). Although paper-and-pencil surveys were not time-stamped, they were handed out before and collected after each lecture, so that responses on paper-and-pencil forms were only possible during the lecture.

The theoretical framework for the data collection originally was Eccles' expectancy-value theory (Eccles et al., 1983), according to which expectancies and values of a task are central motivational forces in students' academic behaviours and learning (Eccles & Wigfield, 2002). They predict academic choices, persistence, and achievement (e.g., Battle & Wigfield, 2003; Cole et al., 2008; Durik et al., 2006). The dataset was also used in previous studies (Dietrich et al., 2017; Dietrich et al., 2019). These previous studies examined associations of situational expectancies and task values with effort (Dietrich et al., 2017), and situational expectancy-value profiles (Dietrich et al., 2019). None of these previous papers analysed the cross-classified data structure in this dataset.

All data and R and Mplus syntaxes for the calculations presented in this article are openly accessible at the Open Science Framework (Moeller et al., 2019).

2.2 Measures

The ESM assessment captured situational task values and expectancies with eight items (see Dietrich et al., 2017 or https://osf.io/qjkmz/). Additionally, situational interest and situational effort were assessed with one item each. The students were instructed to think about the lecture contents of the past couple of minutes and to complete the questionnaire within nine minutes. They were asked "To what extent do the following statements apply to you in the present moment?" and responded on a 4-point Likert scale ranging from 1 = *does not apply* to 4 = *fully applies*.

In the present article, we constructed an averaged composite score, labelled *situational interest* from two items measuring situational interest ("I am interested in these contents") and situational intrinsic value ("I like these contents"). While most ESM studies assess constructs with single items to keep the burden on the participants as low as possible, we opted to assess each construct with multiple items (see Dietrich et al., 2017 or https://osf.io/qjkmz/), based on the idea that the shared variance of multiple indicators is a more reliable indicator of an underlying construct than a single item can be. It could be argued that this approach of using composite scores instead of single items in itself is a contribution to making ESM assessments more objective, as it reduces the risk of confounding a construct of interest with the random and unique influences (unique variance) that a single item captures apart from the construct it is supposed to represent (e.g., a momentary slip in attention causing the respondent to click on a wrong response option, or an idiosyncratic misunderstanding by a given participant of a given item in a given situation). In a cross-classified multilevel model with responses (within level, n = 2,200) nested in both individuals (between-individual level, n = 155) and time points (between-time point level, n = 87), the correlations between the two items of the *situational interest* scale were r = .62 at the within level, r = .93 at the between-individual level, and r = .99 at the between-

time point level.

3. Results and Discussion

3.1 Which concepts of objectivity should be applied to disentangle objective characteristics of situations from participants' subjective perceptions thereof? (RQ1)

When using the term *objectivity*, we assume that there are true characteristics of an object (in this study, a learning situation) that influence individuals' subjective responses in a somewhat systematic way that causes at least partial agreements in the subjective responses of multiple individuals. Skipping the important and millennia-long philosophical discussions about whether or not there is a truth and how to define and understand it (for a summary, see e.g., Glanzberg, 2018), we pragmatically define objectivity here as the (approximate) agreement of all observers about the characteristics of an object, with the object here meaning the learning situation. This here applied concept of objectivity is based on Popper's claim that "the objectivity of scientific statements lies in the fact that they can be inter-subjectively tested" (1934 [2002], p. 22). While Popper refers to scientific statements rather than laymen's implicit concepts, his idea of objectivity as inter-subjective agreement has been extended: Douglas (2011) has named this concept of objectivity the concordant objectivity and defines it as the "simple agreement among multiple observers" (Douglas, 2011, p. 32). This idea of objectivity as the inter-subjective agreement about the truth of an object is reflected in the classical test theory, which typically considers assessments to be objective to the degree that all trained assessors come to the same conclusion about an assessed construct in a given individual or population. It is also reflected in the practice of treating a classes' mean score of averaged individual student responses about perceived teaching quality as a proxy for the objective teaching quality in that classroom (see e.g., Göllner, 2018; Lüdtke et al., 2009). It is important to keep in mind that this is a rather parsimonious concept of objectivity and that many more have been discussed in the social sciences, including Education (e.g., Eisner, 1992; Fisher, 2000).

How can self-reports ever be objective in the sense of concordant objectivity, given that the construct to be assessed (e.g., an emotion) is typically experienced only by the individual who experiences it, which also is the reason why we ask participants to report us their feelings in self-reports? While it is controversially discussed to what degree it is possible to objectively determine how exactly a person feels without relying on this person's subjective self-report (e.g., Barrett, 2018), it is arguably possible and useful to assess the characteristics of a situation that are related to the emotional experiences. For example, since previous research on situational interest suggests that it is largely triggered by objective situation characteristics, such as novel and surprising information being presented, we can expect that multiple individuals agree partially in their situational interest in one learning situation, as long as the information presented to them is equally novel and surprising to each of these individuals.

One possibility of deriving information about objective characteristics of a situation from subjective self-reports is to aggregate the multiple self-reports of different individuals about the same situation. In that sense, the objective assessment of, e.g., a situation's interestingness, would be the agreement of a large enough number of randomly selected individuals about the interest they felt in that situation. In this scenario, we would expect that the group of students tends to report higher situational interest in learning situations presenting students with novel and surprising stimuli, compared to learning situations presenting students with known, unsurprising stimuli, as long as the stimuli in both cases are otherwise similar. The group's average (inter-subjective agreement) of reported interest in a given situation would thus be an indicator of the objective characteristics (interestingness) of that situation.

A limitation of the here-applied definition of objectivity is that the group's mean score in a situation may be sample specific. If we select only the most interested individuals, then their mean interest in a given situation may be high, not because of the situation being novel and surprising, but because of the generally high interest of the group across all situations. Thus, the group mean score in a

given situation may reflect the possible interactions between the group's person characteristics and situation characteristics, rather than indicating the objective situation characteristics alone.

3.2 How does an ESM assessment have to look like for it to capture both the objective situation characteristics and the subjective perceptions thereof? (RQ2)

With the aforementioned definition of objective assessments of situational characteristics through inter-subjective agreements across individual subjective self-reports, we need momentary assessments from multiple individuals in the same situation to aggregate these multiple self-reports to a situation-specific group mean score.

What exactly the term *same situation* means depends on the research question of a given study. For instance, the objective interestingness of a learning situation can be assessed by asking all the students in the same class in the same instant about their current interest in that moment, and then aggregating across these individual responses. If, however, students learn remotely and self-paced with digital learning platforms, then a *situation* in terms of the research question could either be a certain time point (e.g., Tuesdays afternoon, or 24 hours before the final exam), or it could be the individual time point at which each student finishes a given task that is relevant to the research question, or another condition with relevance to the understanding of a digital learning moment. In this article, we use the term *situation* synonymously with a given time point in a given lecture hall in which all students see and hear the same university lecturer talking in the front of the room and are asked at the same time point about their momentary motivation, as described in Dietrich et al. (2017).

In order to assess multiple individuals in the same situation, we need to modify the common design of individually randomly timed survey notifications used in many ESM studies. For our purpose, we need to assess a large enough group of students at the same time, whereas the common ESM schedules typically assesses each student at their own individual random times in order to capture a true random sample of all the experiences that students make during a relevant unit of time, like a school day (e.g., Hektner et al., 2007). If we deviate from such truly random and individual schedules in order to examine students' inter-subjective agreements in one given situation, then the so-collected data may not be a representative sample of everyday life activities.

However, there are many research questions that do not require a representative sample of all everyday life activities. For example, research questions referring to specific school subjects, or specific teachers, or specific lectures, require these contexts to be oversampled to ensure a large enough sample of situational assessments in that chosen context. One challenge in the use of not randomly timed ESM surveys is the risk of systematic context-specific biases in the assessments: The smaller the range of assessed time points or situations, the larger the likelihood of non-random influences. For example, in a truly randomly timed ESM study, we would not expect the results to be influenced by the time of the day, or the students' distractedness during the last minutes of a class when everyone already packs their things to jump up at the first ring of the school bell, or other influences that are particular to a certain timing, because these influences are expected to cancel out. In contrast, if we decide to assess students only in the last five minutes in each class, for instance because teachers are concerned about interruptions and a no-phone policy during lessons, then we cannot rule out that the timing might have biased the responses in a way that a random survey schedule would not have. In order to reduce the risk of contextual biases in the assessments of multiple participants in the same instants, we need to make sure that at least there are no biases concerning the timing of surveys. That means that whatever the time span relevant to the study, no participants should be surveyed only at the beginning or only at the end of that time span. Instead, surveys should be distributed equally across these time spans for all participants.

For that purpose, we have developed an assessment design described below for the study of momentary study motivation in a university course across an entire semester (see also Figures 2 and 3). The weekly 90-minute lessons of that course are split into nine periods of nine minutes each (not ten minutes, because the participants need at least one minute to answer to the last survey in the lecture and would miss that last notification if it occurred when the end-of-lecture-noise and hectic has already

started). To make sure that we have data detailing the motivation across the entire lecture, we assess participants after the first ten minutes, after 19 minutes, after 28 minutes, and so on. We start the first assessment after ten instead of nine minutes, because in the first few minutes, some time tends to get lost on welcoming and waiting for students to calm down. To keep the burden on each individual participant low, each participant is only surveyed three times during the lecture, with a time gap of 27 minutes between each assessment. To assess multiple participants at the same time while still pursuing the aforementioned goals (data across the entire lecture, no participant surveyed more than three times), participants are surveyed in groups, with group A being surveyed after the first 10 minutes, Group B being surveyed 19 minutes into the lecture, Group C being surveyed 28 minutes into the lecture, and then Group A again being surveyed 37 minutes into the lecture, and so on. The same design is then repeated one week later in the same lecture, but with the difference that Group B starts the assessments 10 minutes into the lecture, in order to rotate the survey times across all groups, times, and weeks (Figure 2). Individuals were randomly assigned to groups in a way that ensured a relatively equal sample size of each group.

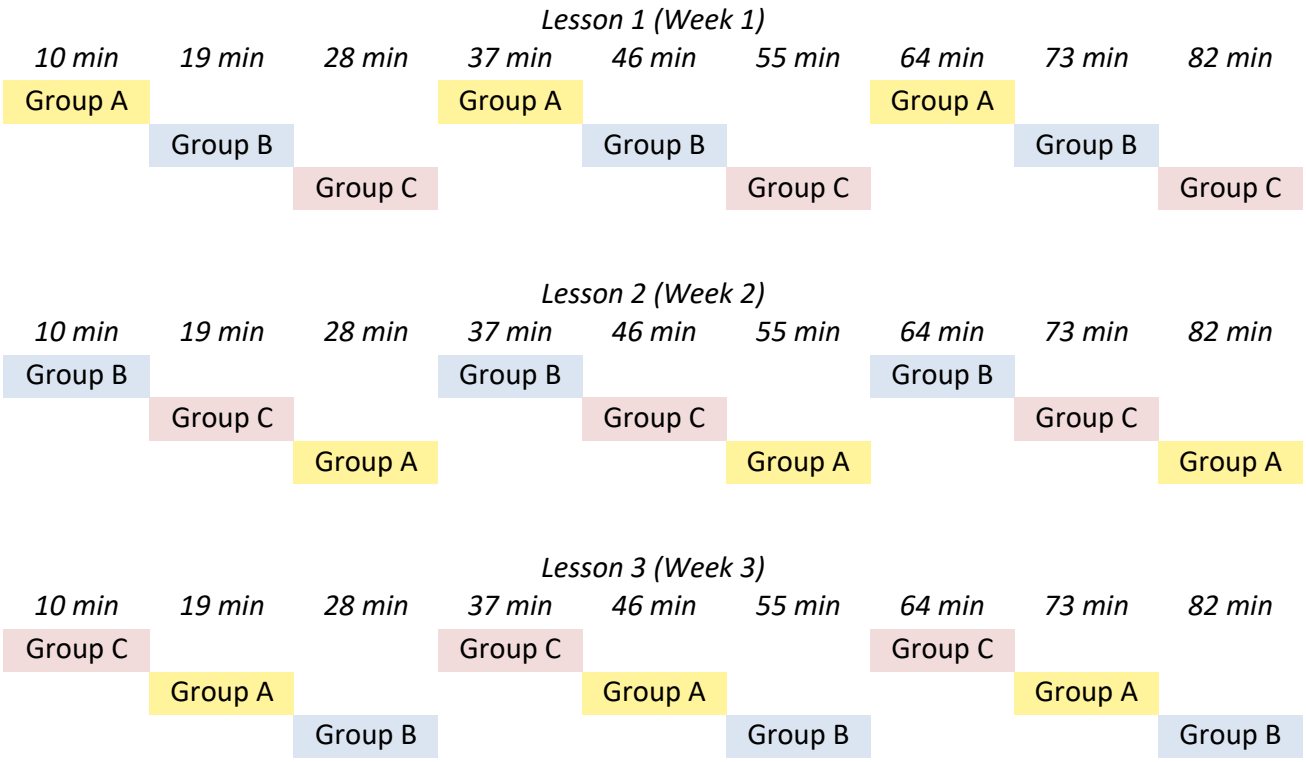


Figure 2. Example signalling schedule in lesson 1, 2, and 3 (to be rotated in following lessons)

3.3 Which analyses are needed to disentangle the objective situation characteristics and the subjective perceptions thereof in data collected with the proposed design? (RQ3)

3.3.1 Analytical strategy 1: Visualising the inter-personal agreement (objective parameter) and the subjective deviation from that agreement (subjective parameter): Jittered violin plots

To start the analyses of the data gathered with the assessment design proposed above, it is recommendable to get an overview of the distribution of the responses at each measurement time point. To explore how much participants agree or individually deviate from the average rating of the interestingness of a learning situation, we suggest examining the inter-individual distribution of interest ratings for every beep in a given lesson with a jittered violin plot (using the R package ggplot2 with the jitter option). Figure 3 shows an example of such a plot.

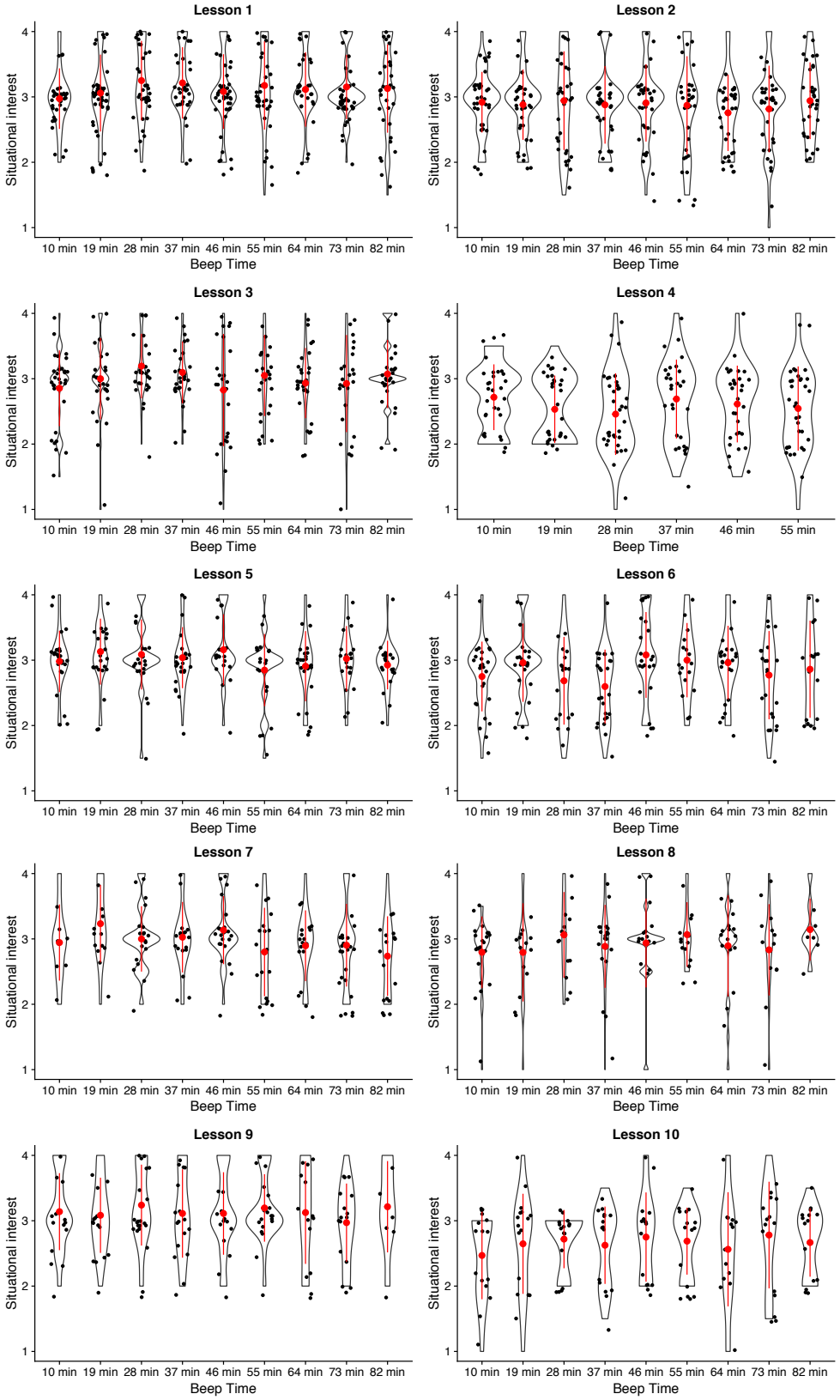


Figure 3. Inter-individual distributions of situational interest for each beep across the ten lessons, visualised in a jittered violin plot (note that lesson 4 ended after 60 minutes, which is why three violins are missing for the last three measurement time points)

The jittered violin plot shows the inter-individual mean score (red dot), the standard deviation (distance between the red dot and one end of the red line), and the inter-individual distribution of interest

ratings (black dots) for each measurement time point / beep in each lesson. The belly of the violin is proportional to the number of individuals who agreed on a rating, with a thick belly meaning that many participants chose this value in their rating of their current interest. Since many observations (dots) would have overlapped on the interest values of 2, 3, and 4, we used the option to *jitter* the dots, which "adds a small amount of random variation on the location of each point" (Wickham et al., n.d.; Wickham, 2016) to prevent overlap and to display the number of the individual responses for each value.

The visual inspection of the violin plots indicates whether the assessed group of students tends to agree, as indicated by violin plots with one clearly distinguished belly (e.g., Figure 3, lesson 1, 10 minutes), or if the responses are randomly distributed across the possible range of values without any particular group agreement, indicated by a flat plot without any belly (e.g., Figure 3, lesson 3, 46 minutes), or if there are two or more distinct groups, each of which agree on their own particular score, indicated by a violin plot with multiple bellies (e.g., Figure 3, lesson 4, 55 minutes). The latter case of a mixed distribution can additionally be examined with statistical tests that tell whether the distribution is unimodal, bi- or multimodal, such as Hartigan's dip test statistic for unimodality versus multimodality (Hartigan & Hartigan, 1985). These tests can be performed with the R package diptest (Maechler, 2016) or the bootstrapping procedure determining the number of modes described in Efron & Tibshirani (1993), which uses the R package bootstrap (Tibshirani & Leisch, 2019). It should be noted that the concept of concordant objectivity applied in this article requires the existence of a unimodal distribution, meaning a high concentration of responses closely around the mean score. If the distribution is bi- or multimodal, then there is no reason to assume that the responses reflect one inter-subjective objectivity, and no reason to claim that the inter-individual mean score was an indicator of such concordant objectivity. Whether or not responses reflect inter-subjectively objectifiable information likely depends on the construct and needs to be determined with the above-mentioned strategies (number of bellies in violin plots and tests for uni- versus multimodality). As Figure 3 shows, the distribution and uni- versus multimodality can also differ from moment to moment, not only from construct to construct. Furthermore, Göllner et al. (2018) have also pointed out that the individual's deviation from the group's mean score does not have to be due to individual rater tendencies or biases, but may represent meaningful information about dyadic experiences. Using the example of teaching quality, the authors argue that different students can make different experiences with the same teacher, so that their deviation from the group mean score may reflect such observable differences in different dyadic experiences. This suggests that not only the group average can serve as an indicator of objective situation characteristics, it is furthermore possible that individual students make experiences that other students don't make, but that other students still would rate the same way if they experienced the same.

3.3.2 Analytical strategy 2: Parameters for the deviation of a subjective rating from the objective group rating: Cross-classified multilevel analyses

After getting a visual impression of the degree of inter-rater agreement on the interestingness of a situation, we might want to calculate parameters quantifying the degree of inter-rater agreement and the degree to which each person at each measurement time point deviates from the group average. In particular, we want to get estimates for (1) the individual and time point-specific deviation from (2) the stable person-specific mean over time, and from the (3) the inter-individual group mean which may differ from situation to situation (i.e. from time point to time point). While component three represents the objective situation characteristic in terms of the characterization the participants can agree on (the average rating of that situation), component one represents the subjective situation-specific deviation from that objective rating, i.e., the subjective element. We would like to control both components 1 and 3 for component two, which represents the stable individual deviation from both the objective situation characteristic and the momentary subjective component due to stable response tendencies of that person (e.g., traits). Such variance decomposition can be done with a cross-classified multilevel analysis (e.g., Beretvas, 2010). This type of statistical model separates the total variance of the scores Y_{it} (the scores of the different individuals *i* at the different time points *t*) in the three aforementioned components:

$$Y_{it} = Y1_{i,t} \text{ (component one, within time point and within individual)} + Y2_{i} \text{ (component two, between individuals)}$$

+ $Y3_t$ (component three, between time points)

The cross-classified multilevel model has the advantage that it allows disentangling the person- and situation-specific deviation (subjective state component) from the group average (objective state component), while accounting for each person's stable tendency to deviate systematically from other individuals across all measurement time points (trait component).

Instead of the more common structure of ESM data, with situations nested only in individuals due to the randomness of the time points, the here proposed assessment design results in time points nested in both individuals ($Y2_i$), and the groups A, B, and C with their respective measurement times ($Y3_t$). Time points are crossed with individuals, because each individual appears only once within each measurement time point. To get reliable estimates about the variance components $Y1_{i,t}$, $Y2_i$, and $Y3_t$, sufficiently big samples of individuals and time points are needed (around $n = 50$ on each of these levels; Chung et al., 2018). The present study design comprises of $n = 155$ individuals and $n = 87$ time points.

We computed the above-described model to separate the total variance [$var(Y_{it}) = .403$] into the three variance components. The biggest portion of the variance pertained to the subjective deviation from both the situational group average and the stable person-specific trait component. This individual, situation-specific component one showed a variance of $var(Y1_{i,t}) = .250$, which equals 62% of the total variance. Second, stable inter-individual differences (traits; component two) accounted for 31% of the variance: $var(Y2_i) = .124$, which means that about one third (31%) of the variance is due to stable person-specific response tendencies that differ between individuals. Finally, the variance of the objective component three was considerably smaller: $var(Y3_t) = .029$, 7% of the total variance. That means in other words that only a small amount of variance was due to changes in the situation-specific group mean score from one moment to another.

3.4 Summary

This study suggested a novel research design that allows to disentangle the objective characteristics of a situation from participants' idiosyncratic momentary subjective deviations from those objective situation characteristics. For example, this novel approach allows disentangling the objective interestingness of a situation from a participant's subjective interest in that moment.

The key of this assessment design is the simultaneous assessment of multiple participants in the same situation / time point, which enables researchers to examine to what degree individuals agree on their ratings of a given construct in that given situation, and to what degree individual participants deviate from that group agreement. We proposed several methods to analyse data assessed with this design and to further examine the role of objective versus subjective components, including jittered violin plots displaying the distribution, means and standard deviations of each measure for each measurement time point, and cross-classified multilevel analysis.

3.5 Practical implications

A ground-breaking advantage of the here proposed assessment design is the fact that it makes feedback to teachers about the *objective* situation characteristics possible. The group agreement indicating the objective interestingness of a situation for example enables teachers to compare their teaching topics and strategies in terms of how they make their class feel. In common momentary assessments in classes, students are typically asked at random time points, implying that for each assessed situation, there is typically only one answer for one individual student. Imagine you were a teacher wanting to know how your new teaching strategy came across to the students, and the researcher tells you: "See, at this time point, ten minutes into your lecture, you introduced the theory of evolution, and Mary reported high boredom and low interest". Would you, as the teacher, conclude that the new teaching strategy failed to raise the students' interests, or would you rather hypothesise about this one student's idiosyncratic reasons for not being interested, or would you remain clueless as to how to interpret this feedback? With the approach of asking multiple students at the same time suggested in this article, it now becomes possible to tell teachers: "See, at this time point, ten minutes into your lecture,

you introduced the theory of evolution, and the average interest reported by your students was high, even though a single student, Mary reported high boredom and low interest". This feedback enables teachers to evaluate the average and the individual perception of their teaching strategies by their students, which we hope will become an important tool for immediate feedback in learning settings.

Imagine for instance that the feedback occurs in real time and the teacher learns that most students are interested but two students are utterly bored. In that case the teacher could offer optional challenging bonus tasks for the two students who might be underwhelmed by the regular classwork. If the entire class is bored, then the teacher could use activating, engaging teaching strategies by trying to cheer up the class with a joke, increasing the task difficulty for everyone, or adding real-life examples allowing students to see the links between the discussed topic and their own interests. As another option, teachers could use the feedback to analyse and revise their teaching strategies and materials after the course or school year has ended. In our studies, we combined the momentary assessments with videos of the lecture, showing both lecturer and slides, allowing us to link the teaching behaviour and materials to the students' momentary motivation. These videos, which are yet to be analysed, are meant to help the teacher (and us researchers) understand which behaviours are most, or least, motivating, and which slides should be modified to foster future students' motivation.

Obviously, collaborations between researchers, and/or software developers and teachers are needed to realise this possibility, unless the researcher and the teacher is the same person (as in the here presented study on motivation in university lectures). The here presented methodological groundwork needed is only the first step in that direction. A next step would require collaborations in which researchers use these methods to identify students' individual needs as well as momentary classroom levels of motivation and emotions. Systematic collaborations of researchers and teachers are needed to provide teachers with the suitable emotion and motivation measures and assessments, and to provide researchers with the real-time data out of real school classrooms. Technology experts are needed for the further development of feasible feedback systems that show the collected data in comprehensible form and real time to students, teachers, and – in the case of underage students – parents. Science communication and more research are needed to find out which form of feedback about the assessed motivation and emotions would be most helpful to students, and teachers.

3.6 Theoretical implications

The combination of the approaches proposed in this article has much potential for the research on motivational heterogeneity of students. The intensive longitudinal data allow for the intra-individual examination of short-term developments (from one measurement time point in a given lesson to the next) and intra-individual long-term development of motivation or emotions (from lesson one to lesson ten). The jittered violin plots can be used to identify particular students as much as they can be used to detect overall trends, like an increase or decrease in the average inter-individual interest from one moment in the lecture to the next.

While common experience sampling method approaches provided no information about a students' deviation from the simultaneously present peer group, the here proposed approach can be used to identify, within any given learning situation, those students who score substantially below the benchmark of interest typical for the simultaneously present peer group. This information potentially makes assessments of learning-related emotions and motivation at the same time more person-specific and more situation-specific. Instead of classifying students as generally less interested than their peers (which a classic ESM approach can do by examining the person-level mean score), our approach enables researchers or educators to say: "Although Mary has a tendency of being less interested than her peers in Math lessons, you really caught her interest with your most recent novel teaching strategy, which brought Mary's interest even above the level of her peers, as you can see in the last two violin plots (where Mary can be marked as a yellow star among the black dots representing her peers)". Thus, we expect that the approach proposed here will make a contribution to personalised learning (e.g., Corno, 2008) and tailored interventions for individual students at the exact times when they are in need of motivational and emotional support. Common experience sampling approaches seem less useful for these purposes, because they leave open whether a given measurement score reflects the individual's

subjective interest or the situations' objective interestingness, or, if both, which of those components to what degree.

By offering techniques to disentangle the idiosyncratic and commonly shared components of motivational self-reports, this article contributed to this special issue's first question ("In what ways do self-report instruments reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use?") and second question ("How does the use of self-report constrain the analytical choices made with that self-report data?"). In sum, our answers to these questions are that self-reports only capture a person's perception but can be aggregated to draw conclusions about the perceptions of a group of persons, their agreements and disagreements, about the characteristics of the (learning) situations they perceive. This article's focus on self-reports of interest in learning settings complements several other articles in this special issue (Chauliac et al., 2020; Fryer et al., 2020; Durik & Jenkins, 2020).

3.7 Limitations of the rotated survey design proposed in this article

One limitation of the design suggested in this article is the fact that the results inherit the problems linked to self-report data, including the fact that self-reports are always to a certain degree idiosyncratic, even when they are averaged or when group agreements are disentangled as a separate source of variance. This implies for example that the group mean score, which we described as the indicator of the objective situation characteristics (e.g., the objective interestingness of a situation) can be sample specific. Imagine if we selected only the most interested individuals for some reason, then their group mean score (objective component) in a given situation will be high, not because the situation is objectively highly interesting but because we only asked the highly interested individuals. Therefore, the objective component of the situational assessments is only objective to the degree to which it can be generalised from the observed sample to a larger population, which is a question for systematic replication studies to examine. If all or many participants in a sample are influenced by similar biases (for instance because we are surveying a group with particularly high social desirability), their agreement (the group average in a given situation) will reflect this joint bias rather than an objective situation characteristic. These are limitations to our definition of objectivity that need to be kept in mind.

Another limitation is the possible diversity of different activities that students who are present in one classroom might be engaged in. For example, in a class taught with a personalised learning approach, different students might be working on different tasks with different instructions. In some personalised learning settings, students in one classroom wear hearing protection to concentrate and receive their individual tasks on technological devices (e.g., tablets) contingent on their prior tasks completed, achievements, and goals. In such settings, it seems unreasonable to assume that the agreement of all raters on, e.g., their current interest, would reflect the objective interestingness of the learning moment, since it is likely that different students were thinking about different tasks when answering.

A third limitation is the requirement of large classes for the design proposed in this article. In order to interpret the distribution and degree of agreement of different raters at any given time point, a reasonably large group is needed. The design proposed in this article was developed for large university lectures, which often involve 200 students or more. Cross-classified models require least 50 students and 50 measurement time points in total (Chung et al., 2018). Per student, at least 10 measurement time points and per measurement time point 10 students responses are needed. However, 10 responses still seem too small of a sample from the standpoint of sampling theory and power considerations, for instance because of the biases that are more likely to affect small samples, compared to larger ones (e.g., Creswell & Guetterman, 2019; Schönbrodt & Perugini, 2013). The purpose and planned analysis should drive the sample size planning, because different approaches require different sample sizes.

In most school classes, it might be less useful to split the class into three groups of responders with different ESM signalling schedules, since many school classes comprise less than 30 students, implying that with the design proposed here, each subgroup at any given time would include no more than ten responses, likely less if school absences, smaller class size, and unwillingness to respond to

ESM signals are taken into account. A possible solution in reasonably large classes might be to signal all students at the same times (e.g., after 25, 50, and 75 minutes of a 90-minute lecture). As a rule of thumb, we recommend to assess all students at the same time if less than 30 students are present, to avoid biases of the group mean score due to outliers. This reduces the number of measurement time points and consequently offers less insight over short-time changes in students' experiences over the course of a lesson, while keeping the burden on each individual student the same (three signals per lesson). If this suggestion is implemented and all students in a school class are surveyed at the same time, then we suggest that the teacher could interrupt the lesson for the duration of the survey to allow students to concentrate on the survey and to avoid that students miss any relevant learning information. To further avoid sample biases, particularly in smaller samples, it might be worthwhile considering to assign matched participants to different groups, so that individuals with similar person characteristics can be found in and compared across all groups. However, it seems unlikely that the needed variety and combinations of person characteristics needed for such matching procedures can be found in small samples such as school classes.

It is furthermore not possible to rule out that nonresponse to ESM signals might be confounded with the constructs being assessed. For example, the most motivated, immersed students might prefer to continue working on their captivating math task and might even miss the ESM signal due to their intense concentration. In some personalised learning classrooms, the headsets that students wear to avoid distractions by their classmates make it difficult to raise their attention to ESM signals, unless these signals come through the same devices their headsets are attached to, which is not always possible. On the other hand, particularly bored and disengaged students might see no point in responding to the ESM surveys. These scenarios of data not missing at random imply that the empirically observed agreement of different students does not necessarily reflect the agreement of all students or the objective situation characteristics, but could itself reflect a biased subsample.

It seems possible that constructs and situation characteristics differ in their potential of being perceived in the same way by different students. It might be easier for students to agree on a question referring to the interestingness of a situation, since situational interest partially depends on observable situation characteristics, such as novelty or surprising information (Hidi & Renninger, 2006), while other constructs may be more person-specific and difficult to observe, such as questions concerning the students' current feeling of competence or frustration. In part, the here proposed design helps detecting and studying such differences between constructs by quantifying the degree to which students agreed in their agreements on different constructs. Nevertheless, it is important to bear in mind that a lack of agreement can have many different sources, either rooted in the construct itself being person-specific, or rooted in individual distractions, individual misunderstandings of items, or individualised instructions.

A general limitation of assessments during ongoing lessons is the risk that the interruptions through surveys, however short, may interfere with the students' attention and learning. This affects all in-the-moment self-reports during classes and consequently most experience sampling method studies conducted in school or university. The here proposed method offers a way out: If it is applied in school, where classes are typically smaller than in university lectures, then all students can be surveyed at the same time and the teacher can stop teaching for the time being. This would still imply an interruption and potential loss of attention, but one that the teacher could afterwards address and try to mitigate, for instance by repeating core messages. In university lectures however, where we suggested to survey different groups of students at different times, it cannot be ruled out that some students might miss an important detail while completing the survey. Teachers who are informed about the survey schedule may want to provide the information presented during survey times afterwards in a format that the student can read and repeat after the lecture, to catch up with any potentially missed information. Future studies should examine whether brief interruptions by ESM surveys interfere with students' learning in school or university. If surveys and teaching occur simultaneously, it cannot be ruled out that the need to split one's attention may impair the accuracy / validity of the students' situational self-report or lead to a selective missing data pattern if students decide not to answer in those learning situations they find most difficult and attention demanding.

There is no guarantee that there be only one group agreement on a given question in a given

situation. There might be multiple subgroups, each with their own mean score / agreed-upon rating. While this would be easy to detect in the violin plot, it poses a limitation to the idea of using the group mean score as the one and only indicator of objective situation characteristics.

If researchers were interested in comparing the groups A, B, and C with each other or make sure that they are comparable, then it would be recommendable to modify the assessment schedule proposed in Figure 2 in a way that assesses two groups simultaneously. With the here-proposed schedule, it would be possible to determine whether group A reported generally higher interest than group B or C, across all situations, by nesting situational assessments (level 1) in individuals (level 2) in groups (level 3), while ignoring the clustering in measurement time points. This procedure would reveal how much variance is due to differences between groups. Alternatively, a multi-group comparison with parameters constraint to be equal (e.g., Asparouhov & Muthén, 2012) could be used to test the assumption that groups were comparable in their mean scores, variances, or other, co-variance-based parameters. However, the procedures suggested in this article do not allow to disentangle the group-specific influence from the situation-specific influence in a given measurement time point, meaning if group A scores particularly high in the 37th minute of the class (see Figure 2), we do not know for sure if group B would have scored the same in a similar situation. If this information is needed for a research questions or application, we recommend to use planned missing data designs that systematically assess multiple (at least 2) groups at a time, in order to be able to compare them (see Enders, 2010). Please note that assessing multiple groups at a time either increases the burden and interruptions for participants, if the schedule is kept the same and the number of surveys is increased for individuals, or it implies fewer measurement time points across the lesson, if the number of individual surveys is kept constant.

Finally, it cannot be ruled out that surveying students' in their learning situation changes the very process we aim to study (e.g., Schmitz & Perels, 2011). This should be kept in mind in all experience sampling method studies surveying students in class, as well in studies using introspective self-reports in general.

3.8 Directions for future method development

We mentioned above that the concept of concordant objectivity employed in this article implies that students assessed in a given situation agree, which in turn implies that their responses should form an uni-modal distribution. However, it is possible that students agree while forming heterogeneous subgroups, leading to a mixture, e.g., bi-modal distribution. For example, the 155 students in our lecture might have consisted of two groups, the ones loving the teacher, and the ones hating the teacher, which might have lead to the bi-modal responses observed in some situations. It is also possible, and apparently was the case in this study, that the form of the distribution of responses varies from moment to moment. It would, for instance, be possible that students form two separate groups in assessing a political statement, with one group of, e.g., conservative students rating the joke as funny and appropriate, and another group of, e.g., liberal, students rating the joke as not funny and inappropriate, or vice versa. Such an instance might cause a temporary bimodal distribution, while all other moments in the same lecture might see a uni-modal distribution as long as no politically connoted jokes are made. In moments in which the distribution is multi-modal, then it would be interesting to find out what caused the distribution. Understanding the reasons and mechanisms behind heterogeneity in responses in given situations is yet to be examined more systematically in future studies. In addition, the variance of the scores at each time point can be small or large, independent of the form of the distribution. For example, even in a study in which all distributions of scores at all measurement time points were uni-modal, the range of scores and the overall variance of scores could be large or small, and could differ from moment to moment. Figure 3 illustrates the size and change in variance between measurement time points in form of the red lines, which represent the standard deviation. Importantly, a mixed distribution (multimodal distribution) suggests multiple groups hiding behind an overall trend, which is highly relevant for personalised learning and person-oriented methods. Thus, apart of quantification of the variance, additional analyses, such as examinations of distributions and cluster/latent profile analyses could complement the search for reasons and mechanisms behind heterogeneity in responses in given situations.

In this study we found that only 7% of the variance was due to changes in the situation-specific

group mean score from one moment to another. While this might seem to suggest that it might not matter so much how a university teacher teaches, we would like to offer alternative interpretations and directions for future research: On one hand, we do not know whether seminars or practical courses at university, which allow for more diverse, hands-on learning experiences than lectures, might have differed more strongly in their average motivation from one moment to another. Our findings only suggest that the lecture examined in this study was relatively consistent in the average interest it elicited from one moment to another (which oscillated around 3 on a scale from 1 = *does not apply* to 4 = *fully applies).* Future studies could examine whether the diversity and distinctiveness of learning tasks in a university course can increase the variance due to differences between changes in the situation-specific group mean score from one moment to another. The fact that the largest proportion of variance (65%) was due to the individual, situation-specific component is a strong argument for personalised learning tools and other instruments that help teachers address the motivational heterogeneity they encounter in their university courses and classrooms. Individual students' momentary motivation deviated much from the average motivation in a given moment in this lecture, and Figure 1 shows that in most moments, there were very interested as well as rather disinterested students present. While heterogeneity and individualised/personalised learning have been addressed increasingly in the literature on learning and instruction in schools (e.g., Banister et al., 2014; Bingham et al., 2018), there is still a need to implement personalised learning procedures in university teaching.

It should be noted that in this article, we do not make use of the longitudinal nature of the ESM data, because that was beyond the article's main scope, which focused on the distinction between subjective and objective components in self-reports. Nevertheless, the here-proposed method also has interesting implications for the longitudinal study of learning and teaching processes. Our method allows to study the following longitudinal questions: How does the construct of choice (here: situational interest) change within a lesson, within a person, over 30 minutes? How does the construct of choice change in one session of a lecture, across individuals, over 9 minutes? How does the construct of choice change from one week to another, over the course of a semester, on average across individuals or within individuals? Reitzle and Dietrich (2019) give an overview of possible longitudinal models that can be used to examine such questions, using the data described in this article and providing corresponding R and Mplus scripts.

While the methods proposed in this article attempt to contribute to further developments of personalised learning and interventions based on momentary assessments, it is important to keep in mind that much more research is needed to get from assessments to valid and helpful interventions. As Bastiaansen et al. (2019) have shown, different teams of researchers can draw very different conclusions about needed interventions from the exact same intensive longitudinal dataset and its intra-individual analyses. Teams of software developers, methodologists, and educators will need to work together to identify valid and effective ways to draw conclusions about individual students' emotional needs for support from data like ours, and about the best ways to deliver appropriate interventions in the appropriate moments.

3.9 Future directions to overcoming the general limitations of self-reports

Because of the general limitations of self-reports, it is important to validate self-report data gained with the here proposed research design by linking them to more objective, observable and behavioural data, such as video-recorded observations of the students' behaviour or the teacher's behaviour, psychophysiological data with relevance to emotions and motivation, such as mobile electrodermal resistance assessments or heart rate variability measures, verifiable information about students' performance (ideally standardised test performance), absenteeism, school dropout and objective information about the students' demographic background, such as their family's household income.

If only self-report assessments are possible due to organisational or other constraints, then different question formats can help to avoid at least the biases typical to rating scales: For example, emotions could be assessed with both open-ended questions ("Please write down here how you currently feel"), which can be linked to rating scales after being automatically analysed with sentiment analyses

tools (e.g., Silge & Robinson, 2017) or manual coding (e.g., Moeller et al., 2018). Researchers and practitioners around the globe work on methods to gather objective information about participants' emotions and motivations. For example, there are studies and companies that retrieve information about people's emotions from their voices (e.g., Krothapalli & Koolagudi, 2013), countless companies and data scientists analyse texts produced by participants for markers of emotions in so-called sentiment analyses (e.g., Altrabsheh et al., 2013), wearable heart rate variability sensors are marketed to researchers and private users with the promise that they will provide objective information about the stress, sleep, recovery, and physical exercise of the wearer (e.g., Firstbeat, 2012). Multiple sensors are integrated to optimise predictions of behaviour and emotions, and machine learning algorithms help integrate all these data, reaching never before seem accuracies in predicting emotions and behaviour (e.g., Carroll et al., 2013).

On the other hand, a recently emerging debate has questioned whether the subjective information about personal experiences provided by self-reports can be entirely replaced by objective measures, since e.g., Barrett (2018) has suggested that even the presumably *objective* markers of emotions are to some extent idiosyncratic. For these reasons, we might have to keep asking people for their self-reports if we really want to know how an individual feels in a given situation, since the subjective evaluation, a crucial part of the emotional experience, is not always captured in the observable and behavioural measures.

For these reasons, we believe that the research design proposed here will remain a useful tool to examine in the future to what degree a given ESM response in a given situation was idiosyncratic and thus a reflection of person-specific characteristics, or in line with the assessments of other students in the same situation. We hope that the indicators of concordant objectivity proposed here can be compared and integrated with other objective measures of emotions and motivation in learning situations in the future, in order to improve predictions of students' learning and behaviour. There is a large array of constructs that could be assessed in line with the here-proposed person-object logic (Figure 1) and schedule for disentangling the subjective and concordant-objective aspects of participants' situational self-reports. Apart from the example of interest discussed throughout this article, the method promises to be insightful also for constructs such as perceived teacher behaviour, students' rating of teaching quality (see e.g., Göllner et al., 2018), or perceived situation or classroom characteristics (e.g., task difficulty, social climate, see e.g., Lüdtke et al., 2009).

Keypoints

- This methodological contribution proposes a new assessment design for experience sampling method data collections that enables researchers to disentangle objective person characteristics from subjective perceptions thereof.
- The proposed design makes it possible to study the development of both subjective and objective parameters across the time span of one weekly lecture and an entire semester, while the burden for each person is kept relatively low with three beeps per lecture.
- Different options for corresponding analyses are proposed, including jittered violin plots for visual inspection, tests for uni- versus multi-modality, and cross-classified multilevel models.
- We discuss implications of the proposed research design for the development of teacher feedback for measures of momentary student emotion and motivation.

Acknowledgments

This research has been supported by a Jacobs Foundation Early Career Research Fellowship awarded to the first author. We thank our reviewers for very thoughtful and appreciative feedback.

References

Altrabsheh, N., Gaber, M. M., & Cocea, M. (2013). SA-E: Sentiment analysis for education. In: R. Neves-Silva, J. Watada, G. Philipps-Wren, L. C. Jain, & R. J. Howlett (Eds.), *Intelligent Decision Technologies (pp. 353 - 362)*, Amsterdam: IOS Press. doi: 10.3233/978-1-61499-264-6-353

Asparouhov, T., & Muthén, B. (2012). Multiple group multilevel analysis. Mplus Web Notes: No. 16. Retrieved March 5, 2020 from https://www.statmodel.com/examples/webnotes/webnote16.pdf

Asparouhov, T. & Muthén, B. (2019). Comparison of models for the analysis of intensive longitudinal data, *Structural Equation Modeling: A Multidisciplinary Journal, 00*: 1–23. https://doi.org/10.1080/10705511.2019.1626733

Banister, S., Reinhart, R., & Ross, C. (2014). Using digital resources to support personalized learning experiences in K-12 classrooms: The evolution of mobile devices as innovations in schools in Northwest Ohio. In M. Searson & M. Ochoa (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2014 (pp. 2715-2721)*. Chesapeake, VA: Association for the Advancement of Computing in Education. Retrieved March 5, 2020 from https://www.learntechlib.org/primary/p/131202/.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F., Boker, S. M., Ceulemans, E., Chen, M., … Bringmann, L. F. (2019, March 21). Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology. *Preprint retrieved on August 24, 2019 from* https://doi.org/10.31234/osf.io/c8vp7

Bingham, A. J., Pane, J. F., Steiner, E. D., & Hamilton, L. S. (2018). Ahead of the curve: Implementation challenges in personalised learning school models. *Educational Policy, 32(3)*, 454 – 489. https://doi.org/10.1177/0895904816637688

Barrett, L. F. (2018). *How emotions are made. The secret life of the brain*. Mariner Books: New York.

Battle, A., & Wigfield, A. (2003). College women's value orientations toward family, career, and graduate school. *Journal of Vocational Behavior*, *62*, 56–75. https://doi.org/10.1016/S0001-8791(02)00037-4

Beretvas, S. N. (2010). Cross-classified and multiple membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis (pp. 313–334)*. New York, NY: Routledge.

Bieg, M., Goetz, T., Wolter, I., & Hall, N. C. (2015). Gender stereotype endorsement differentially predicts girls' and boys' trait-state discrepancy in math anxiety. *Frontiers in Psychology, 6, 1404*. https://doi.org/10.3389/fpsyg.2015.01404

Carroll, E. A., Czerwinski, M., Roseway, A., Kapoor, A., Johns, P., Rowan, K., & Schraefel, M. C. (2013). Food and mood: Just-in-time support for emotional eating. *2013 Humaine Association Conference of Affective Computing and Intelligent Interaction*. Geneva, Switzerland.

Chauliac, M; Catrysse, L. ; Gijbels, D. & Donche V. (2020). It is all in the surv-eye: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*. 8 (3), 26 – 39. https://doi.org/10.14786/flr.v8i3.489

Chung, H., Kim, J., Park, R., & Jean, H. (2018). The impact of sample size in cross-classified multiple membership multilevel models. *Journal of Modern Applied Statistical Methods, 17 (1)*, Article 26. https://doi.org/10.22237/jmasm/1542209860

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*, 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43(3),* 161–173. https://doi.org/10.1080/00461520802178466

Creswell, J. W. & Guetterman, T. C. (2019). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*, 6th edition, Pearson.

Dietrich, J., Viljaranta, J., Moeller, J., & Kracke, B. (2017). Situational expectancies and task values: Associations with students' effort. *Learning and Instruction, 47*, 53–64. https://doi.org/10.1016/j.learninstruc.2016.10.009

Dietrich, J., Moeller, J., Guo, J., Viljaranta, J., & Kracke, B. (2019a). In-the-moment profiles of expectancies, task values, and costs. *Frontiers in Psychology, 10:1662.* https://doi.org/10.3389/fpsyg.2019.01662

Durik, A. M. & Jenkins J. S. (2020). Variability in Certainty of Self-Reported Interest: Implications for Theory and Research. *Frontline Learning Research*. 8 (3) 85-103. https://doi.org/10.14786/flr.v8i3.491

Douglas, H., (2011). Facts, values, and objectivity. In: I. Jarvie & J. Zamora Bonilla (eds.), *The SAGE Handbook of Philosophy of Social Science*, 513–529, London: SAGE Publications.

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology*, *98*, 382–393. https://doi.org/10.1037/0022-0663.98.2.382

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives (pp.74–146)*. San Francisco, CA: Freeman.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, London.

Eisner, E. (1992). Objectivity in educational research. *Curriculum Inquiry, 22(1)*, 9-15. https://doi.org/10.1080/03626784.1992.11075389

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Fahrenberg, J. (1996). Ambulatory assessment: Issues and perspectives. In: Fahrenberg, J. & Myrtek, M. (Eds.). (1996). *Ambulatory Assessment: Computer-assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies (pp. 3 – 20)*. Seattle, WA: Hogrefe and Huber. University of Freiburg i. Br., Germany

Fink, B. (1991). Interest development as structural change in person-object relationships. In: Oppenheimer L., Valsiner J. (eds) *The Origins of Action*. Springer, New York, NY. https://doi.org/10.1007

Firstbeat (2012). Heart beat based recovery analysis for athletic training. *Firstbeat Whitepapers.* Retrieved from: http://www.firstbeat.fi/physiology/white-papers

Fisher W. P. Jr. (2000). Objectivity in psychosocial measurement: what, why, how. *Journal of Outcome Measurement, 4(2)*, 527-563.

Fryer, L. K. & Nakao K. (2020). The Future of Survey Self-report: An experiment contrasting Likert, VAS, Slide, and Swipe touch interfaces. *Frontline Learning Research*, 8 (3),10-25. https://doi.org/10.14786/flr.v8i3.501

Glanzberg, M. (2018). Truth. In: Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Retrieved from https://plato.stanford.edu/archives/fall2018/entries/truth/

Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects

between students and teachers? *Journal of Educational Psychology, 110(5)*, 709–725. https://doi.org/10.1037/edu0000236

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science*, *24*(10), 2079-2087. https://doi.org/10.1177/0956797613486989

Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods, 11*, 87–105. https://doi.org/10.1037/1082-989X.11.1.87

Hartigan, J. A., & Hartigan, P. M. (1985) The dip test of unimodality. *Annals of Statistics, 13*, 70–84.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method. Measuring the quality of everyday life*. Thousand Oaks, CA, US: Sage Publications.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111-127. https://doi.org/10.1207/s15326985ep4102_4

Ketonen, E., Dietrich, J., Moeller, J., Salmela-Aro, K., & Lonka, K. (2018). The influence of autonomous and controlled daily goals on positive and negative emotional states: An experience sampling approach. *Learning and Instruction, 53,* 10-20. https://doi.org/10.1016/j.learninstruc.2017.07.003

Krapp, A. (1998). Entwicklung und Förderung von Interessen im Unterricht [Development and promotion of interest in instruction]. *Psychologie in Erziehung und Unterricht, 45*, 186-203.

Krapp, A. (2002). Structural and dynamic aspects of interest development: theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12(4)*, 383-409. https://doi.org/10.1016/S0959-4752(01)00011-1

Krapp, A., & Fink, B. (1992). The development and function of interests during the critical transition from home to preschool. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development (pp. 397–429)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning and development. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development (pp. 3–25)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Krothapalli, K. S. & Koolagudi, S. G. (2013). *Emotion recognition using speech features*. London: Springer

Lüdtke, O., Robitzsch, A., Trautwein, U., Kunter. M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. Contemporary *Educational Psychology 34*, 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Maechler, M. (2016). *Package 'diptest'. Hartigan's dip test statistic for unimodality – Corrected*. R package. Retrieved March 5, 2020 from https://cran.r-project.org/web/packages/diptest/diptest.pdf

Moeller, J., Dietrich, J., Viljaranta, J., & Kracke, B. (2019). Data, R and Mplus codes for disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Rerieved from https://osf.io/yszvm/*. https://doi.org/10.17605/OSF.IO/YSZVM

Moeller, J., Ivcevic, Z., White, A., & Brackett, M. A. (2018). Mixed emotions: network analyses of intra-individual co-occurrences within and across situations. *Emotion,18(8),* 1106-1121. https://doi.org/10.1037/emo0000419

Popper, K. R. (1934 [2002]), *Logik der Forschung [The Logic of Scientific Discovery*], Berlin: Akademie Verlag.

Prenzel, M., Krapp, A. & Schiefele, H. (1986). Grundzüge einer pädagogischen Interessentheorie [Outline of an educational interest theory]. *Zeitschrift für Pädagogik, 32(2)*, 163-173.

Reitzle, M. & Dietrich, J. (2019). From between-person statstics to within-person dynamics. *Diskurs Kindheits- und Jugendforschung, 3-2019*, 319-339. https://doi.org/10.3224/diskurs.v14i3.06

Schmitz, B. & Perels, F. (2011). Self-monitoring of self-regulation during math homework behaviour using standardized diaries. *Metacognition & Learning, 6*, 255-273. https://doi.org/10.1007/s11409-011-9076-6

Schönbrodt, F. D. & Perugini, M. (2013). At what sampe size do correlations stabilize? *Journal of Research on Personality, 47*, 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology,* 4, 1-32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Silge, J. & Robinson, D. (2017). *Text mining with R: A tidy approach*. Sebastopol, CA: O'Reilly

Takarangi, M. K. T., Garry, M., & Loftus, E. F. (2006). Dear diary, is plastic better than paper? I can't remember: Comment on Green, Rafaeli, Bolger, Shrout, and Reis (2006). *Psychological Methods, 11*, 119 –122. https://doi.org/10.1037/1082-989X.11.1.119

Tibshirani, R. & Leisch, F. (2019). *bootstrap: Functions for the Book "An Introduction to the Bootstrap. R package*. https://cran.r-project.org/web/packages/bootstrap/index.html

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., & Woo, K. (n.d.). Jittered points. Retrieved from: https://ggplot2.tidyverse.org/reference/geom_jitter.html

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No 3 special issue (2020) 85 - 103
ISSN 2295-3159

Variability in Certainty of Self-Reported Interest: Implications for Theory and Research

Amanda M. Durik[a] & Jade S. Jenkins[a]

[a]Northern Illinois University, USA

Article received 15 May 2019 / revised 29 August / accepted 29 August / available online 30 March

Abstract

These studies examined self-reported interest, how level of interest is related to reported certainty of interest, and whether certainty helps to clarify the relationship between interest and behavior. This research borrows from research on attitudes showing that attitude certainty helps to clarify the relationship between attitudes and behavior. A pilot study examined the relationship between self-reported interest and certainty of interest within four disciplines (math, psychology, biology, and astronomy). These relationships were replicated in math (Study 1) and psychology (Study 2), and the relationships between interest and behavior were stronger for those with greater certainty. For domains in which participants had sufficient levels of experience and varied levels of interest, curvilinear relationships were found between level of interest and certainty, showing that certainty is higher among individuals who report more extreme (high or low) levels of interest. Moreover, self-reported interest predicted behavior more strongly for those with more certainty in their responses. Discussion surrounds the theoretical and methodological utility of considering certainty of interest alongside measures of self-reported interest.

Keywords: self-report; certainty; interest

Amanda Durik, Department of Psychology, Northern Illinois University, DeKalb, IL 60115 United States of America, adurik@niu.edu DOI: https://doi.org/10.14786/flr.v8i3.491

1. Background

When scanning a classroom, it is hard not to notice variability in student motivation. The students who are attentive, active, listening, and responding can be differentiated from those who are off task and distracted. Some of this differentiation can be traced to students' varying levels of interest in the domain that is being taught. Interest is an emotional response to particular stimuli that has both cognitive and motivational features (see reviews by Hidi & Renninger, 2006; Krapp, 2002; Prenzel, 1992). It can guide students toward better self-regulation while learning because it helps students focus on content, choose to engage, persevere through challenges, and recall more. Interest is conceptualized as both residing in the person over time (individual interest) as well as varying in the moment in response to stimuli in the immediate situation (situational interest). As such, interest can fluctuate in response to processes operating within the person and stimuli present in the situation.

Individual interest is an enduring tendency to approach and seek learning opportunities in a given domain (Ainley & Ainley, 2011; Deci, 1992; Hidi & Renninger, 2006; Krapp, 2002; Prenzel, 1992; Renninger et al., 1992; Schiefele, 1991; Silvia, 2006). The domain is associated with affective involvement and meaningfulness (Schiefele, 1991), which are enhanced by the acquisition and use of knowledge (Hidi & Renninger, 2006; Renninger, 2000; Prenzel, 1992). A hallmark of individual interest is willingness to reengage in the domain over time (Hidi & Renninger, 2006).

The measurement of interest has been given considerable attention. Options of how to measure interest often include self-report scales and interviews (see discussions by Krapp & Prenzel, 2011; Renninger & Hidi, 2011). Within those, decisions about which features of interest should be captured vary. For example, within the collection of self-report scales that are available, some include feelings associated with interacting with domain content, the meaning or value associated with the domain, and/or the perceived knowledge or actual knowledge that individuals have stored in the domain (see review by Renninger & Hidi, 2011).

Most conceptualizations of interest reflect the idea that interest can change in response to experience and exposure to content, and we argue that research on interest must inherently take a developmental perspective. This leads to challenges in measurement because any measure is a snapshot of a person's interest at a particular moment. Moreover, in the case of closed-ended self-report measures of interest, a given response is that person's best attempt at quantifying their level of interest at that time.

These issues also manifest in the capacity (or lack of, in some cases) for self-report measures of interest to predict behavioral manifestations of interest. Behavioral measures of interest often assess whether and how people choose to engage with domain content. Not surprisingly, measures of self-reported interest often positively predict behavior, such as free-choice behaviors observed in the laboratory, course taking, retrospective reports of behavior, and intentions to behave in the future (Ainley, Hidi, & Berndoff, 2002; Harackiewicz et al., 2008; Renninger, 1990; Simpkins et al., 2006; Wijnia et al., 2014). That said, although correlations between self-reported interest and behavior are often present, they may not be as strong as one might expect.

Other areas of research have carefully considered why self-reported and behavioral data are not always strongly associated, and social psychologists who study attitudes began working on this issue within their own area in the 1970s. This work toward understanding the relationship between self-reported attitudes and behavior has led to greater clarity surrounding the nature of and research on attitudes. Similarly, careful attention to the relationship between self-reported interest and behavior may also help clarify the construct of interest. The challenges encountered by attitude researchers in assessing self-reported attitudes are likely similar to many of the challenges encountered by interest researchers assessing self-reported interest. In both cases, participants are asked to evaluate their responses to a particular class of stimuli or ideas. Although it is simple enough for an individual to provide a response on a scale, this deceivingly simple response is the outcome of much more complex processes.

It should be noted, however, that we do not argue that interest and attitudes are the same. Interest assumes an active process on the part of the individual that propels them toward knowledge acquisition,

elaboration, or growth (Deci, 1992) whereas an attitude does not necessarily trigger this process and may actually do the opposite. For example, a person who holds a positive attitude toward recycling would believe that recycling is good. A person who holds a negative attitude about recycling would believe that recycling is bad, possibly a waste. Whereas the person with a positive attitude may be more open to learning about recycling than the person with a negative attitude, the attitude itself does not motivate learning. In contrast, a person with an interest in recycling would be expected to have learning goals related to recycling (e.g., a desire to learn about the processes related to recycling, which materials can be recycled, and why they should be recycled).

Attitude researchers have addressed the issue of attitude-behavior consistency in several ways. For example, one approach recognized that other environmental variables such as norms and the opportunity and ability to engage in the behavior were also critical (Ajzen, 1991). Another approach recognized that attitudes predicted behavior more strongly when people focused on only one side of an attitude (e.g., for or against; Glasman & Albarracin, 2006). Finally, another approach identified that other attitude features contributed to attitude strength, and increased the relationship between attitudes and behavior (Krosnick & Petty, 1995).

Borrowing from this last approach, the current research centers on the idea that individuals vary in the extent to which they are certain of their attitudes. Certainty refers to the extent to which individuals are confident in their assessment of an attitude as clear and correct (Rucker et al., 2014). Certain attitudes are held more strongly and are less likely to change (Krosnick & Petty, 1995; Pomerantz et al., 1995). Moreover, participants who reported greater certainty of their attitudes, which was measured separately from the attitudes themselves, were more likely to behave in ways that were consistent with their attitudes (e.g., Bizer et al., 2006; Fazio & Zanna, 1978; Glasman & Albarracıin, 2006; Tormala, 2016; Tormala & Rucker, 2007). As an example, although participants may report varying levels of attitudes toward recycling, those who have more certain and positive attitudes toward recycling would be more likely to actually recycle.

Just as individuals can report less or more certainty of their attitudes, we theorized that some participants would be more certain of their self-reported interest and others less so. We reasoned that if attitude researchers were able to clarify the relationship between attitudes and behaviors by considering attitude certainty, it was worthwhile to attempt the same for individual interest.

This approach, compared to some of the other approaches explored within the attitude literature, was selected because it preserved the assumption that research on interest must consider development. Interest changes and people may become more certain of their interest over time. As such, not only might the inclusion of certainty clarify the relationship between interest and behavior, but it may also provide insight into how participants' awareness of interest changes.

Specifically, certainty of interest may prove useful in gaining insight into the nature of interest and how individuals come to recognize their interests. Drawing again from the attitude literature, the extent to which individuals become more confident or certain of their attitude is related to the extent and valence of prior experiences and the amount of careful thought put toward the object of the attitude (Berger, 1992; Bizer et al., 2006; Fazio & Zanna, 1978; Glasman & Albarracin, 2006; Jonas et al., 1997; Krishnan & Smith, 1998; Prislin et al., 1998). Prior experience and careful thought toward the object of an attitude have been found to increase certainty, which then contributes to stability of attitudes. The current research examines variability in certainty as a starting point in determining whether similar processes may be operating for interest as they are for attitudes.

The first aim of the current research is to examine certainty of interest and how certainty varies with levels of self-reported interest. The second aim was to test whether individuals who are more certain behave in ways that align more closely to their interests.

2. Pilot Study

The purpose of the pilot study was to explore the patterns of association between self-reported interest and certainty of interest among different domains (math, biology, astronomy, and psychology). These domains were chosen because it was expected that participants would vary in their prior experience with each. The pilot sample was composed entirely of advanced psychology students. As such, this population was anticipated to have varying levels of exposure to math and biology (due to compulsory education), high exposure to psychology (as their program of study), and low exposure to astronomy (neither compulsory nor inherently linked to their program of study). This anticipated variability in experience with the different domains may have implications for certainty, and create meaningful comparisons across the domains.

We tested whether certainty and interest would be related in a linear or curvilinear fashion, and were especially interested in a curvilinear relationship such that participants who reported more extreme levels of interest (either low or high) may also be more certain of their interest. This pattern was found in prior research on attitudes revealing a curvilinear relationship between certainty and willingness to advocate for an attitudinal position (Cheatham & Tormala, 2017). Moreover, if the relationship was linear, the redundancy in interest and certainty may undermine the utility of considering certainty of interest as separate from interest.

2.1 Method

Design

This was a correlational study using a within-participants design in which participants answered questions about their interest and certainty of interest in four domains.

2.1.1 Participants

The participants were 21 undergraduate students at a mid-sized university in the Midwestern United States. They were all in an upper-level psychology course that students typically complete their last year of undergraduate study. They completed the questionnaire in exchange for extra credit.

2.1.2 Measures and procedure

Participants responded to a 4-page, paper-and-pencil survey, in which questions for each domain (math, biology, astronomy, and psychology) were presented on different pages. The order in which participants responded about the different domains was counterbalanced across participants.

Participants responded to items assessing interest, certainty, and the number of college courses they had completed in the domain as well as other items that are not central to the current research. Interest was measured with 6 items that were adapted from those used in prior research and capture both feeling and meaning/value aspects of interest (Durik & Harackiewicz, 2007). The scale included "I find ___ interesting," "___is fascinating to me," "I find ___enjoyable," "___is a boring subject," "___just doesn't appeal to me," and "I think ___is a meaningful discipline." wherein the blank spaces were replaced with the domain name. Participants rated each item from 1 (*Strongly disagree*) to 7 (*Strongly agree*). Cronbach alphas for interest were .94 (math), .90 (biology), .77 (psychology), and .81 (astronomy). The lower reliability observed for psychology is likely due to restriction of range because all participants were highly interested in psychology, which is known to constrain estimates of reliability (Nunnally & Bernstein, 1994).

Certainty was assessed with one item, "How sure are you of your attitudes about ___?" and rated from 1 (*Not at all*) to 5 (*Very much*). Although only a single item was used, a similar approach has been taken in prior research (Gross et al., 1995). Participants were also asked, "How many college courses have you taken in ___?" with response options ranging from zero to greater than eleven. This item was included in order to examine whether certainty was related to participants' prior exposure to the domain.

2.2 Results and discussion

The data for this study (and both subsequent studies) were analyzed using multiple regression analysis conducted in SPSS Version 25.0. For each domain, certainty was designated as the criterion variable. The measure of interest in each domain was standardized and a squared term was calculated by multiplying the standardized measure by itself. These two variables, the standardized measure of interest and its square, were entered into the regression simultaneously to predict certainty of interest for that domain. For each effect, squared semi-partial correlations are provided as measures of effect size. These denote the portion of total variability in the outcome variable that is uniquely accounted for by a given predictor.

2.2.1 Math

The analysis predicting certainty of math interest revealed a negative average relationship of interest, $t(18) = -2.48$, $p = .02$, $B = -0.42$, $sr^2 = .23$, and a positive quadratic relationship, $t(18) = 2.54$, $p = .02$, $B = 0.38$, $sr^2 = .24$. The top left panel of Figure 1 depicts the relationship, showing that certainty was higher for those reporting either low or high levels of interest, and lower for those reporting more moderate levels of interest.

2.2.2. Biology

The analysis predicting certainty of biology interest revealed no relationship of interest, $t(18) = 0.68$, $p = .51$, $B = 0.12$, $sr^2 = .02$, but similar to math, yielded a positive quadratic relationship, $t(18) = 3.12$, $p < .01$, $B = 0.65$, $sr^2 = .35$. The bottom left panel of Figure 1 depicts the relationship. Similar to what was observed in math, participants who reported lower or higher levels of interest in biology also reported greater certainty, compared with those who reported more moderate levels of interest.

2.2.3 Astronomy

The model used to predict certainty of astronomy interest revealed a different pattern. Neither a linear relationship, $t(18) = 0.68$, $p = .50$, $B = 0.10$, $sr^2 = .02$, nor a quadratic relationship, $t(18) = -0.18$, $p = .86$, $B = -0.02$, $sr^2 < .01$, emerged (see top right panel of Figure 1).

2.2.4 Psychology

The regression predicting certainty of psychology interest yielded a positive linear relationship of interest, $t(18) = 3.34$, $p < .01$, $B = 0.24$, $sr^2 = .12$, as well as a negative quadratic relationship, $t(18) = -2.56$, $p = .02$, $B = -0.19$, $sr^2 = .07$. The bottom right panel of Figure 1 reveals a different quadratic relationship than was observed for math and biology. Among this sample of upper-level psychology students, interest appears to be positively related to certainty, and then levels off at the highest levels of interest and certainty.

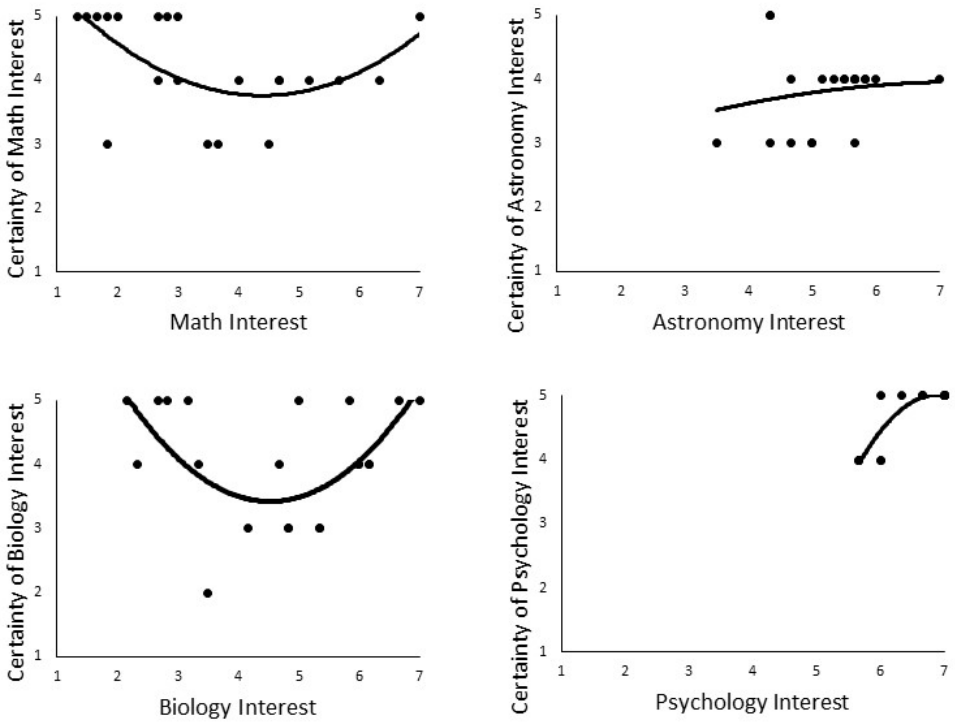


Figure 1. Curvilinear relationships tested between interest and certainty for each domain in the Pilot Study. Math and biology (left panels) revealed statistically significant ($p < .05$) positive quadratic relationships, psychology (bottom right) revealed a significant negative quadratic relationship, and astronomy revealed no relationship (top right).

Overall these analyses of the relationships among level of interest and certainty reveals several things. First, the relationship between interest and certainty varies considerably by domain, such that negative quadratic relationships are observed in this sample for math and biology, a positive quadratic relationship was observed in this sample for psychology, and no relationship was observed for astronomy. These varied relationships are likely due to both the amount of experience these participants have with each domain, as well as their high interest in psychology due to the fact that participants were sampled from an upper-level psychology class. It seems likely that the quadratic relationships observed in math and biology are representative of most domains in which individuals have sufficient exposure to the domain to report their interest, and assuming that the sample includes the full range of interest. In most cases when participants are asked to report their interest, they have sufficient experience from which to draw conclusions about and be aware of their level of interest (either high or low) in the domain. In contrast, we interpreted the absence of relationships found in the astronomy domain as likely due to participants having had little prior exposure to astronomy. Lack of experience likely limited both their level of certainty as well as their extremity of interest in the domain. Finally, the results for psychology revealed a different quadratic pattern from the other three and reflects this sample's high certainty and high interest in psychology.

To examine whether certainty did covary with participants' prior experience, a final analysis was performed in which reports of certainty and the number of college courses reported was aggregated across participants. A correlation was calculated between the average number of courses students reported for each domain and the average level of certainty for each of the four domains. The correlation of the aggregated measures revealed a strong and positive association, $r(2) = .99$, $p < .01$, suggesting that the number of reported courses among participants in this sample was strongly and positively associated with level of certainty. One

could imagine that a more thorough measure of prior experience, including courses taken in secondary school or informal learning opportunities, would add further insight into this relationship.

Although the sample size for this pilot study was extremely small and only included participants who were highly interested in psychology, the results were promising enough to explore further. Studies 1 and 2 were designed to examine these relationships more in depth within two domains, math and psychology, among participants drawn from a more general population.

3. Study 1

Study 1 was designed to replicate the results observed in the pilot study in the domain of math. Given that certainty of interest is likely to increase as individuals have more exposure to domains, and that students are exposed to years of compulsory math in primary and secondary school, we expected the negative quadratic relationship between interest and certainty that was observed in the pilot study to also emerge in Study 1. The second purpose of Study 1 was to test the relationship between self-reported interest and behavior for those with lower or higher certainty. We hypothesized that self-reported interest in the domain would predict behavior in the domain positively and more strongly if individuals were more versus less certain of their interest.

3.1 Method

3.1.1 Design

This was a correlational study that took place in a laboratory context. Participants' math interest, certainty of math interest, and math-related behaviors were assessed in a single session.

3.1.2 Participants

The participants were 138 undergraduate students (54% women) completing an introductory psychology course at a mid-sized university in the Midwest United States. They participated in the study for partial course credit. The sample included participants who reported their race or ethnicity as African American (24%), Hispanic (16%), Asian (9%), Caucasian (48%), or as another, unlisted category (3%).

3.1.3 Measures and procedure

Participants were invited into the lab individually and completed the measures using MediaLab software (Jarvis, 2004). First, participants reported their interest in math using five items adapted from prior research (Harackiewicz et al. 2008), including "I've always been fascinated by mathematics," "I'm really excited about learning mathematics," "I'm really looking forward to learning more about mathematics," "I think mathematics is an important discipline," and "I think mathematics is important for me to know." Participants responded to each item from 1 (*Strongly disagree*) to 7 (*Strongly agree*). The internal consistency of the scale items was strong (Cronbach's alpha = .90).

Participants responded to 6 items designed to assess their certainty of interest (Krosnick et al., 1993). These items were, "How CERTAIN are you of your feelings toward mathematics?", "How SURE are you that your opinion of mathematics is correct?", "How FIRM are your opinions of mathematics?", "How EASILY could your opinion of mathematics be changed?" (reversed), "How DEFINITE are your views of mathematics?", and "How CONVINCED are you of your views of mathematics?" from 1 (*Not at all* ___) to 7 (*Very* ___), in which the blank restated the word in the question that was presented in capital letters. The reversed item was omitted because it decreased the internal consistency of the scale, which left 5 items (final Cronbach's alpha = .91).

Participants also had the opportunity to report their engagement in math-related behaviors. Behavioral indicators are often influenced by many factors in a given situation so the three behavioral indicators were combined into a composite after being standardized. One set of items asked participants to reflect on the past two years and indicate whether or not they had voluntarily chosen to engage in 12 activities related to math, including "I have surfed a website about mathematics in my spare time," "I have voluntarily discussed topics related to mathematics with friends or family," "I have chosen to join a club related to mathematics," and "I have spent free time reading a magazine article about mathematics." The number of behaviors indicated were summed for each participant.

Two additional behavioral measures occurred during the session. Participants were given a set of 10 math-related topics (e.g., "polygons and figures," "pi," "statistics," "quadratic equations") and asked to mark any about which they would like to receive more information via email (and to provide their contact information to do so). Participants were also given the opportunity to watch any of 5 short video clips about math-related topics (e.g., pi, mental math techniques, square roots). The number of topics marked and the number of videos watched were summed and each served as an additional measure of behavior. The standardized scores for all three measures were averaged to obtain the behavior composite (Cronbach's alpha = .69).

3.2 Results and Discussion

An exploratory factor analysis using oblique rotation was conducted to explore whether the measure of certainty was different from that of interest. Two eigenvalues over 1 emerged and the pattern matrix showed two factors with fairly simple structure. Each item had a loading of at least .74 on its expected factor and no loading over .07 on the unexpected factor.

The next analysis focused on testing the relationship between level of interest in math and certainty of math interest. As was done in the pilot study, this was achieved by conducting a multiple regression analysis in which certainty served as the criterion variable, and a standardized measure of interest as well as its square served as the two predictors. This analysis revealed both a positive relationship of interest, $t(135) = 4.68$, $p < .01$, $B = 0.38$, $sr^2 = .12$. as well as a positive quadratic relationship, $t(135) = 6.02$, $p < .01$, $B = 0.47$, $sr^2 = .20$. Comparable to the pattern that emerged in the pilot study with regard to math, those with either lower or higher levels of interest in math also reported more certainty. In contrast, those who reported a moderate amount of interest in math reported lower certainty (see Figure 2).

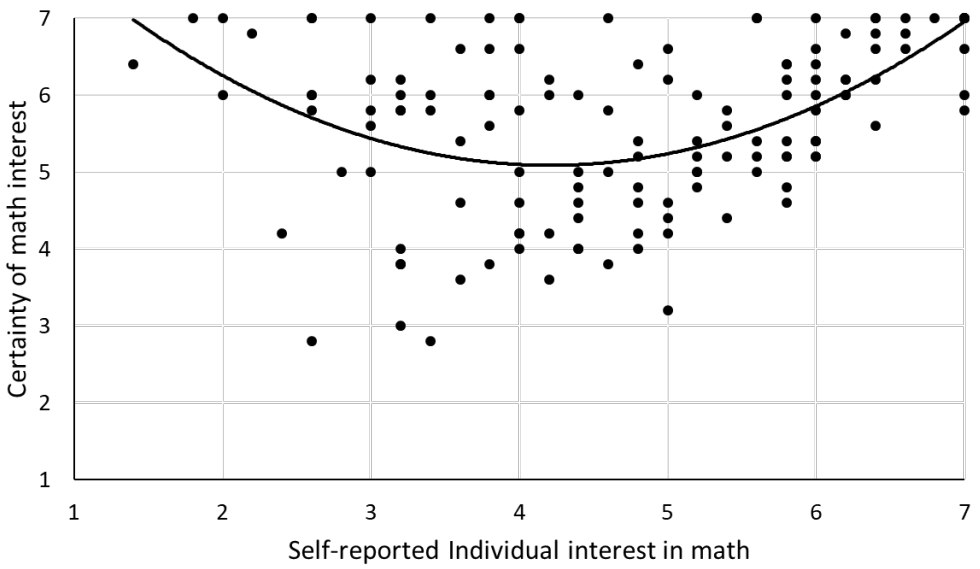


Figure 2. Curvilinear relationship observed between interest in math and certainty in Study 1.

The second analysis focused on whether the relationship between math interest and behavior would be positive and stronger for participants who reported greater certainty. To this end, a second regression analysis was conducted. The criterion variable was the composite measure of behavior and the three predictors included standardized measures of math interest, certainty of math interest, and their product. The analysis yielded a strong positive relationship of interest, $t(134) = 4.67$, $p < .01$, $B = 0.32$, $sr^2 = .12$, and the predicted interaction, $t(134) = 2.57$, $p = .01$, $B = 0.17$, $sr^2 = .04$, indicating that the relationship between interest and behavior varied depending on level of certainty. Certainty was not a significant predictor. Simple slope analyses were conducted to examine the relationship between interest and behavior separately for participants reporting certainty that was one standard deviation above and below the mean. The relationship between math interest and behavior was significant and positive for those with high certainty, $t(134) = 7.12$, $p < .01$, $B = 0.49$, but not significant for those with low certainty, $t(134) = 1.28$, $p = .20$, $B = 0.15$ (see Figure 3).

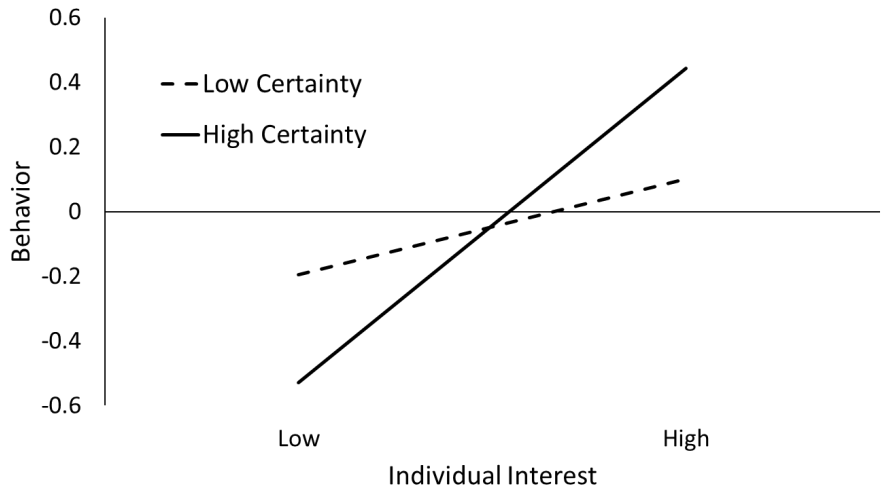


Figure 3. Interaction depicting different relationships between math interest and behavior for those reporting lower (one *SD* below the mean) and higher (one *SD* above the mean) certainty in Study 1.

These data replicate the pattern observed in the pilot study between math interest and certainty, and also showed that individuals who are more certain of their interest are more likely to behave in ways that are consistent with their levels of interest. In other words, those who are more certain of having lower interest are less likely to engage whereas those who are more certain of having higher interest are especially likely to engage. An interesting picture begins to emerge with regard to participants who are less certain. What is learned from Study 1 is that these participants' behavior is not as closely associated with their interest.

4. Study 2

Study 1 offered additional evidence that individuals' self-reported interest and certainty vary in a curvilinear way, and that participants with higher certainty are more likely to behave in ways that are consistent with their level of reported interest. Study 2 was designed to test this again but in a different domain, psychology. Psychology was chosen for two reasons. First, the pilot study showed that interest in psychology and certainty showed a different relationship (a negative quadratic relationship) in contrast to the positive quadratic relationship observed in Study 1 as well as the two other domains examined in the pilot study. We suspected that the observed relationship between interest and certainty for psychology found in the pilot study was due to the sample being composed entirely of highly interested psychology students nearing the completion of their degree. That said, it could instead be due to the domain itself. We wanted to examine the relationship between interest and certainty in psychology, but with a more general sample.

Second, psychology offered the possibility of testing the ideas related to interest and certainty with regard to individual interest as well as situational interest, given that the students were completing introductory psychology at the time that the data were collected. Whereas individual interest is thought of as an enduring person characteristic, situational interest refers to interest that is triggered by cues in the environment (Hidi & Renninger, 2006; Mitchell, 1993; Schraw & Lehman, 2001). Although individual interest and situational interest are often highly correlated, we reasoned that certainty may function differently for these two types of interest. Among students enrolled in an introductory psychology class, it was possible to examine overall individual interest in psychology as well as situational interest in the class, and compare how each measure of interest predicted behavior when taking into account level of certainty. We did not make hypotheses about whether one measure of interest (individual or situational) would predict behavior more strongly than the other.

4.1 Method

4.1.1 Design

This was a correlational study in which participants' interest in psychology, certainty of their interest, and behaviors related to the domain of psychology were assessed.

4.1.2 Participants

The participants included 142 undergraduate students (55% women) completing an introductory psychology course at a mid-sized university in the Midwest United States. They participated in the study for partial course credit. Participants in the sample reported their race or ethnicity as African American (20%), Hispanic (10%), both African American and Hispanic (1%), Asian (3%), Caucasian (65%), or as another, unlisted category (1%). One person did not respond to the question about race/ethnicity.

4.1.3 Measures and procedure

There were two measures of interest, both individual and situational. To report individual interest, participants rated whether "Psychology is…" "interesting," "stimulating," "boring" (reversed), "engaging," "meaningful," "worthless" (reversed), and "useful" from 1 (*Not at all*) to 7 (*Very much*; Schiefele, 1990). Situational interest was measured with eight items reflecting the students' level of interest in their introduction

to psychology course (e.g., "What we are learning in psychology class this semester is fascinating to me" and "We are learning valuable things in psychology class this semester"; Linnenbrink-Garcia et al., 2010). Participants responded to each item on a scale from 1 (*Strongly disagree*) to 7 (*Strongly agree*). The Cronbach's alphas for individual and situational interest were .88 and .95, respectively.

Participants reported their certainty immediately following both measures of interest. The items that measured certainty were the same as in Study 1 and were general in that they did not specify whether participants should report certainty of individual or situational interest. Cronbach's alpha for the certainty measure was .88.

At the end of the survey, reports of behavior were measured with two types of items, and the two types were standardized and combined into a composite in the same way as in Study 1. In parallel with Study 1, participants were asked whether or not they had engaged in 11 psychology-related behaviors in the past two years (e.g., "I have chosen to join a club related to psychology," "I have spent free time reading a magazine article about psychology"). Participants were also given the option of indicating whether they would like to receive information about various psychology-related topics, and if so, to mark their topic choices and provide their email address so this information could be sent. Fifteen topics were listed, designed to capture broad areas of psychology (e.g., "How the brain works," "Mental illness," "How memories form," and "Stereotyping and prejudice"). The total behaviors indicated and the total number of topics selected were both standardized and then averaged to form the composite measure of behavior. Given that there were only two types of behaviors, Cronbach's alpha for this measure was modest equaling .48, which may attenuate the relationships observed between interest and behavior. In contrast to Study 1, the option to offer participants videos to watch was not possible because Study 2 used pencil-and-paper surveys.

4.2 Results and Discussion

4.2.1 Individual interest

As in Study 1, an exploratory factor analysis was conducted on the certainty and interest items in order to evaluate their structure. The structure was not as clean as in Study 1, due to the two interest measures (individual and situational) having items that shared variability. This is not terribly surprising given the similarity in the constructs, methods of measurement, and timing. That said, the certainty items tended to load together and separately from the interest items, again attesting to the uniqueness of the certainty measure as a complement to typical measures of interest. We proceeded with the two separate measures of interest given their conceptual distinction but also recognize that the similarity in their measurement may hinder the ability to see predictive differences across the two measures.

As in Study 1, the first analysis was designed to test the relationship between level of interest and certainty, which was then followed by an analysis to test the relationship between self-reported interest and behavior, with the addition of certainty as a moderator.

A multiple regression model was tested using certainty of interest as the criterion variable, and interest and its square as the predictors. As in Study 1, the measures of interest were standardized prior to calculating the squared term. Replicating the relationship observed in Study 1 with the domain of math, individual interest in psychology had both a linear, $t(139) = 5.68$, $p < .01$, $B = 0.53$, $sr^2 = .18$, and quadratic relationship with certainty, $t(139) = 4.00$, $p < .01$, $B = 0.22$, $sr^2 = .09$. Similar to math, and unlike the relationship observed in the pilot study, certainty was higher for those with lower or higher interest in psychology, and lower for those who reported more moderate interest.

Next, to test whether interest was a stronger predictor of behavior for those with more certain interest, a multiple regression model was tested in which the behavior composite was the criterion variable and the three predictors were the standardized composite measure of individual interest, the standardized measure of certainty, and their product. This analyses revealed a positive relationship between interest and behavior, $t(138) = 5.09$, $p < .01$, $B = 0.34$, $sr^2 = .10$, as well as an interaction, $t(138) = 1.99$, $p < .05$, $B = 0.12$, $sr^2 = .02$ (see

Figure 4). Simple slope tests were conducted to examine the relationship between interest and behavior for those scoring one standard deviation below and above the mean of certainty. These analyses revealed that the relationship between individual interest and behavior was significant and positive for both, but stronger for those with higher certainty, $t(138) = 6.04$, $p < .01$, $B = 0.46$, than for those with lower certainty, $t(138) = 2.17$, $p = .03$, $B = 0.22$.

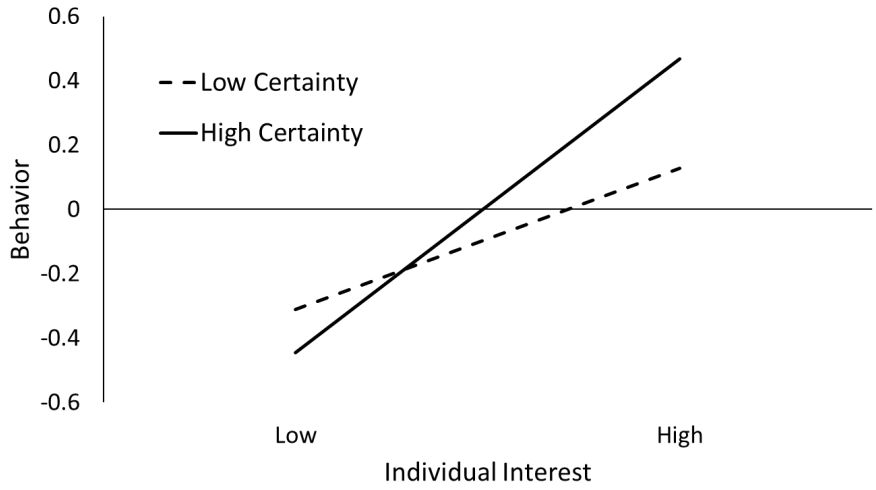


Figure 4. Interaction depicting different relationships between psychology interest and behavior for those reporting lower (one *SD* below the mean) and higher (one *SD* above the mean) certainty in Study 2.

These relationships replicate the patterns that were observed in Study 1. Furthermore, the pattern between interest in psychology and certainty that was observed in the pilot study seems to have been due to the sampling of participants for the pilot study and not to the domain.

4.2.2 Situational interest

The next set of analyses were parallel to those described for individual interest, but instead focused on situational interest.

When situational interest in the psychology class was used to predict certainty, both a linear relationship, $t(139) = 5.48$, $p < .01$, $B = 0.48$, $sr^2 = .17$, and a quadratic relationship, $t(139) = 3.17$, $p = .02$, $B = 0.21$, $sr^2 = .06$, emerged. Participants who reported more extreme levels of situational interest also reported greater certainty whereas those who reported more moderate levels of situational interest reported less certainty.

When situational interest, certainty, and the interaction between them were used to predict the composite measure of behavior, both situational interest, $t(138) = 2.71$, $p < .01$, $B = 0.19$, $sr^2 = .05$, and their interaction were significant, $t(138) = 2.29$, $p = .02$, $B = 0.15$, $sr^2 = .03$. Simple slopes were tested and showed that situational interest positively predicted behavior for those with higher certainty, $t(138) = 4.10$, $p < .01$, $B = 0.34$, but did not predict behavior for those with lower certainty, $t(138) = 0.40$, $p = .69$, $B = 0.04$.

The results from both individual and situational interest demonstrate that certainty may be useful in better predicting behavior with both types of interest. If anything, the interaction was slightly more pronounced for situational than for individual interest, although this difference was not tested directly. One explanation is that the salience of an ongoing situation may be a vivid motivator of behavior toward (or away from) domain content. However, this pattern is also tied to the nature of the situation assessed here. Given that the situation that defined situational interest in this study was in reference to a semester-long, introductory class, situational interest as well as the behavior were in reference to a fairly broad definition of the field in general. Other

situations that are more narrow (e.g., focused on a particular topic) may not predict behavior as strongly, especially if the behavioral opportunity is more broad or more narrow than the experience of situational interest.

5. General Discussion

This research lays out some of the complexity inherent in asking individuals to self-report their interest and recognizes that these self-reports may be more or less certain. For domains in which individuals have varied amounts of experiences and sufficient prior exposure (e.g., math, psychology, and biology), there appears to be a curvilinear relationship between self-reported interest and self-reported certainty. Individuals who reported more extreme levels of interest (either high or low) tended to report being more certain of their interest. In other words, when individuals were quite interested in a domain, they were sure of the presence of their interest; when individuals were quite disinterested in a domain, they were also sure of their lack of interest. Moreover, as predicted, level of interest was a stronger predictor of behavior when certainty was high than when certainty was low. It is not clear from these data whether participants who are less certain (and more moderate in their interest) have truly neutral beliefs about the domain or if they actually have both positive and negative experiences, which are best captured as neutral on this bidirectional scale.

5.1 Promises and Pitfalls of Self-Report Measures

Certainty of interest highlights both a promise and a pitfall of using self-report measures to assess interest. This research speaks to two of the three main questions guiding the compilation of this special issue. First, this research relates to the complexity of interpreting self-report data. On the one hand, if a sample has high certainty of their interest, then self-report measures may be easier to interpret. The certainty with which people report their interest may support processes that strengthen associations between interest and various behavioral outcomes.

In considering this promise, however, it is also critical to consider the pitfall; when people are not certain, they will still provide a response but that response may not be rooted in as much experience, which could make interpretation difficult. If participants are asked to provide reports of interest in domains in which they have little experience, their reports will be ill-informed. This challenge adds to the challenge of inattentiveness, examined by Iaconelli and Wolters (2020). The interpretation of self-report data will likely be murky not only when people are responding inattentively, but also when they are attentive but do not have sufficient self-knowledge or experience with which to provide a meaningful response.

Although the current work centers on the certainty with which individuals can identify their interests, certainty may extend to other types of reports as well, such as certainty in metacognition and strategy use. Other authors in this issue (e.g., Rogiers et al., 2020; van Halem et al., 2020), report how self-reported measures link to trace measures that occur during task engagement, and to subsequent behavior. These relationships may be stronger to the extent that participants are more certain of their self-reported meta-cognition and strategy use.

Second, this research also highlights constraints of self-report data that have implications for methodology. Self-reported domain interest may be less valid when interest is in early phases of development (e.g., Renninger & Hidi, 2011), and this research suggests that certainty may capture an important element. It may take months and years for individuals to collect information about their response to a given domain in order to have certainty in their level of interest (either high or low). Presumably, the development of this certainty will emerge as individuals have experiences with a domain and then think back to them retrospectively (see Dinsmore et al., this issue for a discussion of retrospective processes). If interest must be measured among a sample with limited certainty of interest, one approach may be to provide supports for them

to know how to respond (e.g., definitions of the domain, a particular experience to reflect on) in order to provide a more valid measure, or to assess interest multiple times during a task (Moeller et al., 2020) rather than relying on a global assessment of domain interest.

Alternatively, researchers may want to consider certainty when selecting domains to study and identifying appropriate samples within those domains. It is noteworthy that the sample in the pilot study had relatively little certainty about their interest in astronomy and very strong certainty about their interest in psychology. Given the results of Studies 1 and 2, this might also have implications for behavior. If the purpose of a research study is to use interest to predict behavior, then it may be wise to select a domain in which participants have considerable experience in the domain (i.e., math), because interest may predict behavior if the sample as a whole has more experience with the domain. That said, it is also important to have variability in the sample so that it represents the full range of interest, with which to predict behavior. The pilot study included a sample composed entirely of students in their last year of studying psychology. Although this sample reported very high certainty, restriction of range in their interest is likely to have limited any observed association between level of interest and behavior, had behavior been assessed.

5.2 Implications for Theory

The observed relationship between interest and certainty may be related to participants' experiences with domain content, and captures the changing aspect of interest within a developmental trajectory. Data from the pilot study showed that the number of classes students took in each domain positively predicted certainty in a linear fashion. Prior experience also varied along with the different observed relationships between interest and certainty across the domains. No reliable relationship was detected in the domain of astronomy, likely because participants had such limited experiences in the domain, and the opposite quadratic relationship was detected in the domain of psychology, presumably because students had extremely high interest and certainty. Although these fluctuations are consistent with research on attitudes showing that direct experience contributes to attitude certainty (see review by Glasman & Albarracin, 2006), the nuances of how this occurs is important to consider for understanding how individuals come to recognize and be aware of their interest.

One possibility is that the extremity of the emotion prompts awareness of the experience and contributes to certainty. Individuals who have relatively intense experiences in a domain—experiences that are either highly interesting or highly distasteful—may come to realize their level of interest and have greater certainty (Dutta et al., 1972). In contrast, those who have more mixed or vague reactions may be less aware of their emotional reaction to the domain content, which then leaves them with less certainty of their interest.

Another possibility is that the clarity or vividness of individuals' memories contributes to certainty. For example, those who have had multiple, semester-long courses in the domain may have more vivid memories of learning in that domain, which may contribute to greater certainty. Along these lines, it may be fruitful to bring research on autobiographical memory into research on interest in order to better understand how individuals recall prior experiences that may inform their report of interest. Research suggests that the valence of how an event ends has a disproportionate influence on how the event is remembered (Kahneman et al., 1993). Research on interest may benefit by building on this foundation from the memory literature in order to better understand not only how interest develops, but whether the timing of experiences impacts how people recall the events that they come to see as foundational to their perception of interest in the domain.

5.3 Implications for Research

Certainty may also provide a handle for predicting whose interest may be more or less altered by new experiences with a domain. For example, individuals who are less certain of their interest may be more responsive to situational variables designed to affect interest. Less certain individuals may be more open to collecting information through their experiences with domain content and updating their level of interest. If so, situational enhancements designed to foster interest may be more effective for low versus high certain

individuals. As such, it may be beneficial to assess certainty of interest in research testing interventions designed to foster interest. The effects of a situational intervention may be positive for a subset of individuals (i.e., those with lower certainty), but this effect may be masked by others who are more certain and therefore less likely to change their level of interest.

In a sense, if one thinks of situational interest as being analogous to attitudes in the face of persuasive attempts, then the attitudes literature provides some hints of this possibility as well. For example, Tormala (2016) has noted how curiosity is itself a form of "interested uncertainty" which, in some situations, can aid persuasion. Specifically, Kupor and Tormala (2015, Study 3) reason that curiosity motivates more thorough processing of the persuasive message and increases the message's impact. This may suggest that a similar process could explain differential impact of situational enhancements that are designed to foster interest. In general, future directions focused on underlying cognitive and affective processes are warranted and would very much help illuminate how interest may or may not change in response to new experiences.

5.4 Differences Between Attitudes and Interest

Although this research was initiated because of similarity between interest and attitudes, these constructs are not the same, which has implications for how they emerge and function. For example, certainty in the attitude literature has been divided into correctness and clarity (Petrocelli et al., 2007). Whereas correctness reflects the veraciousness of an attitude in an absolute sense (e.g., efforts to slow global warming is the correct perspective), clarity is the extent to which individuals are confident that they know their own stance (i.e., I know my attitude about efforts to slow global warming). Applied to interest, clarity is more relevant than correctness if one assumes a stance of relativism, in which interest in one domain is not more valuable or correct than interest in any other domain.

A difference between the attitude and interest literatures also emerges when one considers what is being evaluated when self-assessing an attitude versus self-assessing level of interest. Attitude objects are typically outside the person and the question is whether the person agrees with or disagrees with the attitude object. This is in contrast to interest in which the person assesses *how* they interact with the domain. When people evaluate their interest, they evaluate their personal experiences with it and how they respond to the domain. As such, assessments of an attitude object may be more about the object and less about how the person interacts with the object. For interest, in contrast, the assessment is about one's own response to the domain.

5.5 Limitations

Finally, although positive relationships between interest and behavior were observed in the domains of math and psychology, the direction of causality is not clear in the present correlational studies. The theoretical model describing the relationship between interest and behavior typically places interest as the motivation (i.e., cause) of behavior; however, it is also worth pointing out that behavior may influence self-reports of interest as well as certainty. Research on attitudes has addressed several of these processes (see review by Olson & Stone, 2005). When individuals are unsure of their attitudes, they may reflect on their past behavior as information that can be used to inform their attitude (e.g., Bem, 1967). For example, when asked about global warming, individuals may scan their memories for events in which they chose to engage in pro-environment activities or not. These memories may lead them to decide on a particular self-reported attitude. A similar process may operate with regard to interest, especially when individuals have less certainty of their interest. For example, when participants in the pilot study were asked about their interest in astronomy, the domain in which they had taken the fewest classes, they may have tried to recall relevant memories. Those who could generate more positive memories are likely to have rated their interest higher than those who could generate fewer memories, or negative memories.

It is also worthwhile to consider how behaviors may affect self-reported interest ratings, which can also explain the observed relationships between interest and behavior found in this set of studies. In Studies 1

and 2, participants first reported their interest, then their certainty of interest, and finally responded to behavioral opportunities. It is possible that participants felt pressure to behave in a way that was consistent with their initial reports, which may have strengthened the observed results (Olson & Stone, 2005). Specifically, those who had just reported low or high levels of interest may have felt internal pressure to behave in ways that were consistent with their reports, and this may have been especially strong for those who reported greater certainty. The present research does not address this possibility, but opens up a line of research in which these ideas could be explored.

Finally, specific features of the studies reported here also warrant caution in drawing broad conclusions. The pilot study examined prior exposure to various domains by collecting information on students' course experiences. However, experiences in high school courses and extracurricular activities may have also provided opportunities for exposure. Inclusion of all these experiences could help paint a fuller picture of the relationship between domain exposure and certainty. Furthermore, these studies involve a very narrow sample of individuals—namely, undergraduate students at a single university who are taking a psychology course. For example, it is healthy to question whether similar patterns would be observed among a sample of older adults (e.g., who may have more experience and greater certainty) or younger children (i.e., who may have even less experience than the sample tested in the current research). Moreover, these psychology students may have been especially sensitive to environment cues or demand characteristics related to behaving in ways that are more consistent with their stated interest. This tendency would be exaggerated if participants felt that the desired response involved giving higher interest ratings and engaging in more behaviors. These questions cannot be answered with the current data but could be tested in the future.

5.6 Concluding Thoughts

In summary, when interest is self-reported, the report reflects a synthesis of what individuals have available to them at the moment of measurement, and these reports are more certain for some participants than for others. This variation in certainty may provide a lens for better understanding how interest develops and how it is internalized and becomes known to individuals. As with any type of measure, it is critical to interpret the data in light of the assumptions, capacities, limitations, and processes that are relevant at the time of measurement.

Keypoints

- Self-reported interest varies in certainty across individuals
- Those with greater certainty self-report more extreme interest (high or low)
- Self-reported interest and behaviour correlate more strongly for those with greater certaint

References

Ainley, M., & Ainley, J. (2011). A cultural perspective on the structure of student interest in science. *International Journal of Science Education, 33,* 51-71. https://doi.org/10.1080/09500693.2010.518640

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*, 545-561. https://doi.org/10.1037//0022-0663.94.3.545

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes, 50,* 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183-200. https://doi.org/10.1037/h0024835

Berger, I. E. (1999). The influence of advertising frequency on attitude-behavior consistency: A memory based analysis. *Journal of Social Behavior & Personality, 14,* 547–568.

Bizer, G. Y., Tormala, Z. L., Rucker, D. D., & Petty, R. E. (2006). Memory-based versus on-line processing: Implications for attitude strength. *Journal of Experimental Social Psychology, 42,* 646–653. https://doi.org/10.1016/j.jesp.2005.09.002

Cheatham, L. B., & Tormala, Z. L. (2017). The curvilinear relationship between attitude certainty and attitudinal advocacy. *Personality and Social Psychology Bulletin, 43,* 3-16. https://doi.org/10.1177/0146167216673349

Deci, E. L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K. A. Renninger, S. Hidi, & A. Krapp, A. (Eds.), *The role of interest in learning and development* (pp. 43-70). Hillsdale, NJ: Lawrence Erlbaum.

Dutta, S., Kanungo, R. N; & Freibergs, V. (1972). Retention of affective material: Effects of intensity of affect on retrieval. *Journal of Personality and Social Psychology*, 23, 64–80. https://doi.org/10.1037/h0032790

Fazio, R. H., & Zanna, M. P. (1978). Attitudinal qualities relating to the strength of the attitude– behavior relationship. *Journal of Experimental Social Psychology, 14,* 398–408. https://doi.org/10.1016/0022-1031(78)90035-5

Glasman, L. R., & Albarracin, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological Bulletin, 132*, 778-822. https://doi.org/10.1037/0033-2909.132.5.778

Gross, S. R., Holtz, R., & Miller, N. (1995). Attitude certainty. In R. E. Petty & J. A. Krosnick (Eds), *Attitude strength: Antecedents and consequences* (pp.215-245). Mahwah, NJ: Lawrence Erlbaum.

Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest and performance. *Journal of Educational Psychology, 100(1),* 105-122. https://doi.org/10.1037/0022-0663.100.1.105

Hidi, S. & Renninger, K.A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111-127. https://doi.org/10.1207/s15326985ep4102_4

Iaconelli, R. & Wolters C.A. (2020). Insufficient Effort Responding in Surveys Assessing Self-Regulated Learning: Nuisance or Fatal Flaw? *Frontline Learning Research*. 8 (3) *104 – 125.* https://doi.org/10.14786/flr.v8i3.521

Jarvis, W. B. G. (2004). MediaLab [Computer software]. New York: Empirisoft.

Jonas, K., Diehl, M., & Broemer, P. (1997). Effects of attitudinal ambivalence on information processing and attitude-intention consistency. *Journal of Experimental Social Psychology, 33,* 190–210. https://doi.org/10.1006/jesp.1996.1317

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*. *4,* 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Krapp, A. (2002). An educational-psychological theory of interest and its relation to SDT. In. E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 405-427). Rochester, NY: University of Rochester Press.

Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods, and findings. *International Journal of Science Education, 33,* 27-50. https://doi.org/10.1080/09500693.2010.518645

Krishnan, H. S., & Smith, R. E. (1998). The relative endurance of attitudes, confidence and attitude-behavior consistency: The role of information source and delay. *Journal of Consumer Psychology, 7,* 273–298. https://doi.org/10.1207/s15327663jcp0703_03

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology, 65,* 1132-1151. https://doi.org/ 10.1037/0022-3514.65.6.1132

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1-24). Mahwah, NJ: Lawrence Erlbaum.

Kupor, D. M., & Tormala, Z. L. (2015). Persuasion, interrupted: The effect of momentary interruptions on message processing and persuasion. *Journal of Consumer Research*, 42, 300–315. https://doi.org/ 10.1093/jcr/ucv018

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70,* 647-671. https://doi.org/10.1177/0013164409355699

Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, *85*, 424–436. https://doi.org/10.1037/0022-0663.85.3.424

Moeller, J., Viljaranta, J., Kracke, B., & Dietrich, J. (2020). Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Frontline Learning Research*. 8 (3) 63-84. https://doi.org/10.14786/flr.v8i3.529

Nunnally, J. C., & Bernstein, I. A. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Olson, J. M., & Stone, J. (2005). The influence of behavior on attitudes. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 223-271). Mahwah, NJ: Lawrence Erlbaum.

Petrocelli J. V., Tormala, Z. L., & Rucker, D. D. (2007). Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of Personality and Social Psychology, 92,* 30-41. https://doi.org/10.1037/0022-3514.92.1.30

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69,* 408-419. https://doi.org/10.1037/0022-3514.69.3.408

Prislin, R., Wood, W., & Pool, G. J. (1998). Structural consistency and the deduction of novel from existing attitudes. *Journal of Experimental Social Psychology, 34,* 66–89. https://doi.org/10.1006/jesp.1997.1343

Renninger, K. A. (1990). Children's play interests, representation, and activity. In R. Fivush & K. Hudson (Eds.), *Knowing and remembering in young children* (pp. 127-165). New York: Cambridge University Press.

Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone and J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 373-404). San Diego, CA: Academic Press, Inc.

Renninger, K. A., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist, 46,* 168-184. https://doi.org/10.1080/00461520.2011.587723

Renninger, K. A., Hidi, S., & Krapp, A. (1992). *The role of interest in learning and development.* Hillsdale, NJ: Lawrence Erlbaum.

Renninger, K. A., & Su, S. (2012). Interest and its development. In R. M. Ryan (Ed.), *The Oxford handbook of motivation* (pp. 167-187). Oxford: Oxford University.

Rogiers, A., Merchie, E., & Van Keer H. (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research*. 8(3), 40–62. https://doi.org/10.14786/flr.v8i3.527

Rucker, D. D., Tormala, Z. L., Petty, R. E., & Briñol, P. (2014). Consumer conviction and commitment: An appraisal-based framework for attitude certainty. *Journal of Consumer Psychology*, *24*(1), 119-136. https://doi.org/10.1016/j.jcps.2013.07.001

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist, 26*, 299-323. https://doi.org/10.1080/00461520.1991.9653136

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading, 3,* 257-279. https://doi.org/10.1207/s1532799xssr0303_4

Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13*, 23-52. https://doi.org/10.1023/A:1009004801455

Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. *Emotion, 5,* 89–102. https://doi.org/10.1037/1528-3542.5.1.89

Silvia, P. J. (2006). *Exploring the psychology of interest.* New York: Oxford University Press.

Silvia, P.J. (2008). Appraisal components and emotion traits: Examining the appraisal basis of trait curiosity. *Cognition and Emotion, 22*, 94–113. https://doi.org/10.1080/02699930701298481

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology, 42*, 70-83. https://doi.org/10.1037/0012-1649.42.1.70

Tormala, Z. L. (2016). The role of certainty (and uncertainty) in attitudes and persuasion. *Current Opinion in Psychology*, *10*, 6-11. https://doi.org/10.1016/j.copsyc.2015.10.017

Tormala, Z. L., & Rucker, D. D. (2007). Attitude certainty: A review of past findings and emerging perspectives. *Social and Personality Psychology Compass*, *1*, 469-492. https://doi.org/10.1111/j.1751-9004.2007.00025.x

Van Halem, N., van Klaveren, C., Drachsler H., Schmitz, M., & Cornelisz, I. (2020). Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data. *Frontline Learning Research*. 8 (3) 140-163. https://doi.org/10.14786/flr.v8i3.497

Wijnia, L., Loyens, S. M. M., Derous, E., & Schmidt, H. G. (2014). Do students' topic interest and tutors' instructional style matter in problem-based learning? *Journal of Educational Psychology, 106,* 919-933. https://doi.org/10.1037/a0037119

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 104 - 125
ISSN 2295-3159

Insufficient Effort Responding in Surveys Assessing Self-Regulated Learning: Nuisance or Fatal Flaw?

Ryan Iaconelli & Christopher A. Wolters [1]

[1] Dennis Learning Center, Department of Educational Studies, The Ohio State University, USA

Article received 18 June 2019 / Article revised 9 October / Accepted 30 October / Available online 30 March

Abstract

Despite concerns about their validity, self-report surveys remain the primary data collection method in the research of self-regulated learning (SRL). To address some of these concerns, we took a data set comprised of college students' self-reported beliefs and behaviours related to SRL, assessed across three surveys, and examined it for instances of a specific threat to validity, insufficient effort responding (IER; Huang, et al., 2012). Using four validated indicators of IER, we found the rate of IER to vary between 12-16%. Critically, while we found that students characterised as inattentive and attentive differed in some basic descriptive statistics, the inclusion of inattentive students within the data set did not alter more substantial inferences or conclusions drawn from the data. This study provides the first direct examination of the impact of respondents' attention on the validity of SRL data generated from self-report surveys.

Keywords: insufficient effort responding; self-regulated learning; self-report; validity

Address correspondence to Ryan Iaconelli, Dennis Learning Center, Department of Educational Studies, The Ohio State University, 1640 Neil Ave, Columbus, OH, 43201 Email: iaconelli.1@osu.edu DOI: https://doi.org/10.14786/flr.v8i3.521

1. Introduction

1.1 Use of SRS in Studying Self-Regulated Learning

Researchers have used models of self-regulated learning (SRL) to understand engagement, learning, and achievement in academic contexts from preschool through college (Perry, et al.,, 2018; Pintrich & Zusho, 2007; Usher & Schunk, 2018; Winne & Hadwin, 2008). Models of SRL posit that students can plan, monitor, control, and reflect upon their own thoughts, behaviors, and motivation related to their learning (Panadero, 2017). Engagement in SRL requires that students feel both motivated and efficacious to enact these sub-processes (Pajares, 2007; Pintrich & Zusho, 2007). Efforts to use SRL as a basis for developing instructional policies and practices designed to improve students' academic success is an accepted goal among educators (Cleary & Zimmerman, 2004; Dignath & Buttner, 2008; Schunk & Zimmerman, 1998). Achieving critical goals with regard to these two efforts is in no small part dependent upon the availability of sound methods for assessing SRL (Winnie & Perry, 2000; Wolters & Won, 2018).

The need for sound assessment of SRL has spawned the development of many methods (Azevedo, et al., 2018; Winne & Perry, 2000; Zimmerman, 2008). For instance, observing students' behaviours within the classroom, recording traces of their thinking or behaviour when engaged in academic tasks, and reports by teachers or parents have all been used to assess students' SRL (Biswas, Baker, & Paquette, 2018; Cleary & Callan, 2018). Despite the promising increase in the diversity of assessments, self-report surveys (SRS) remain the most common method used to assess SRL (Winne & Perry, 2000). We use the term SRS to describe any type of questionnaire or survey in which respondents are presented with a question or statement and asked to provide a response, either retrospectively or concurrently, based on their own beliefs, attitudes, or behaviours. Although they offer many advantages (Butler, 2002; McCardle & Hadwin, 2015; Wolters & Won, 2018), criticisms of SRS, including fundamental questions regarding the validity of the data they produce (Karabenick & Zusho, 2011; Schellings & van Hout-Wolters, 2011; Winne & Jamieson-Noel, 2003).

Developments in the manner under which SRS can be completed, such as the increased use of unsupervised internet-based administrations (e.g. Qualtrics, REDcap, Amazon Mechanical Turk), raises concerns about the evidence for validity from data produced when students complete online SRS for educational research. Responding appropriately to items on an SRS is a function of a complex, multi-step process. To authentically respond to an item, the respondent must read and understand the item, search their memory for relevant information, integrate any activated memories into a coherent answer, match this answer to one of the available response options, and finally decide whether to select that or some other response option (Duckworth & Yeager, 2015). At any point in this process, lack of motivated engagement or inattention on the part of the respondent is a threat to the validity of individual items as well as the overall data produced. Hence, it is important to carefully evaluate potential threats to the validity of the data produced from these SRS (Wolters & Won, 2018). We address this need by evaluating college students' responses to three online SRS designed to assess factors associated with their SRL, motivation, and academic success for issues related to inattentive responding.

1.2 Insufficient Effort Responding

Threats to validity that result from respondents' lack of cognitive effort, inattention, or motivation when completing surveys have been examined under several different names, including careless responding (Meade & Craig, 2012), insufficient effort responding (IER; Huang, et al., 2012), and low-quality data (DeSimone & Harms, 2017). We adopt IER as our preferred term and it is defined as "a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instruction, correctly interpret item content, and provide accurate responses" (Huang et al., 2012, p. 100). In other words, IER occurs when a respondent does not provide the necessary cognitive effort required to go through the multi-step process needed to produce data that

appropriately represents the underlying construct. The IER framework and the methods used to identify it encompass physical, cognitive, and motivational disengagement from the survey, all of which threaten its' validity.

IER may occur for many reasons. For instance, a respondent may be fatigued, distracted by their surroundings, or motivated to complete the SRS as fast as possible and without effort, either because they are forced to take the survey or because they are taking it solely for some promised compensation, like money or extra credit in a course (Johnson, 2005). Mischievous responding, in which a respondent purposefully provides systematically invalid responses (e.g. selecting the same response for all items) can be considered a form of IER because, even though they are providing some effort, it is not directed at interpreting and responding to items appropriately. Socially desirable responding and other unintentional biases that may alter a person's response patterns do not fall under the IER umbrella because these respondents are working to read, understand, and answer items appropriately.

Researchers have appreciated the threats to validity represented by the underlying causes of IER for some time (Beach, 1989; Nichols, et al., 1989). It was not until recently though that researchers began to investigate more vigorously the propensity and importance of IER as a detriment to validity. Two key findings have emerged from this work. One, the exact level of IER within a data set varies from survey to survey and sample to sample, but generally about 10% of respondents are identified as having engaged in some form of IER (Meade & Craig, 2012; Maniaci & Rogge, 2014). Two, while this proportion may not seem noteworthy, the inclusion of even a small percentage of inattentive responders, as low as five percent, can cause spurious relationships among otherwise uncorrelated measures to become significant, mask otherwise significant relationships between variables, lead to inflated mean-level scores on latent constructs, and alter effect size differences between groups (Huang, et al., 2015; McKibben & Silva, 2017).

This line of research raises serious concerns for those who rely on SRS as a primary means of data collection, SRL researchers included. There are reasons however to question the extent to which this previous research is applicable to SRL, primary among them is that most of this research has derived from assessments of personality and other more stable traits. The more dynamic and changeable nature of motivation and SRL constructs, compared to personality constructs, may result in different manifestations of IER. Additionally, much of the research on IER has utilized samples of undergraduates drawn from participant pools (e.g. Dunn, et al., 2016; Huang et al., 2012; Meade & Craig, 2012) that presumably had little intrinsic motivation for completing the SRS. In contrast, our sample consists of students who anticipated using the results of their SRS for meaningful diagnostic purposes as part of a course assignment. Thus, in addition to extending IER research into a new field, we also extend this research into a presumably more motivated sample, which should produce different manifestations of IER.

1.2.1 Methods for Identifying IER

To identify instances of IER, researchers can call upon both proactive and reactive indicators. Proactive methods are based on the inclusion of "check items" throughout an SRS (Huang et al., 2012; Meade & Craig, 2012; Huang et al., 2015; Maniaci & Rogge, 2014; Bowling et al., 2016; Dunn et al., 2016) that provide a way to determine if respondents are reading each item carefully. Some check items direct respondents to a certain answer (e.g. "Mark strongly agree for this item"), while others are statements that no respondent should agree with (e.g. "My birthday is February 30"). Finally, some researchers (e.g. Huang et al., 2012) have simply asked respondents how much effort they put forth when completing items. Although proactive measures have proven useful, the focus of this study was on the use of reactive methods to detect IER.

Reactive methods refer to a variety of post-hoc statistical analyses used to identify IER. With the increased use of technology to administer surveys, using total survey response time, which programs like Qualtrics record automatically, has become an easy and effective means of assessing IER (Huang et al., 2012; Meade & Craig, 2012; Maniaci & Rogge, 2014; Bowling et al., 2016; Dunn et al., 2016). The basic assumption of this approach is that those who spend very little time completing the survey are

not fully engaged in the various cognitive processes necessary to respond to items as intended. For instance, respondents may skim items rather than carefully read them or may quickly select a response without deliberate recollection and reasonable consideration of the events that should inform their response. Hence, survey completion times that are extremely short are highly suggestive of IER.

A family of statistical analyses designed to assess the consistency of one's responses are another set of common reactive methods for identifying IER. As one example, even-odd consistency (Huang et al., 2012; Meade & Craig, 2012; Johnson, 2005) is a measure of individual reliability that is generated by dividing a survey into two parts (traditionally based on odd and even-numbered items) and calculating the correlation between the two parts. The underlying assumption of this method is that alternate items from unidimensional scales should be strongly correlated. Another common individual reliability indicator is referred to as psychometric synonyms/antonyms (Curran, 2016). This indicator is computed by identifying the items with the strongest bivariate correlations within a data set and comparing individual respondents' correlations on these items. Semantic synonyms/antonyms are used in the same manner, except that the items used in the calculation of this index are decided on an *a priori* basis, based on item content. For each of these indices, respondents who exhibit unusually weak correlations for the set of relevant pairs of items are thought to have engaged in IER. This conclusion is based on the assumption that the atypical correlations are a function of not reading items carefully enough, answering items randomly, or utilizing another response strategy that falls substantially short of full engagement in the response process, a process which provides increased evidence for the validity of the data (Duckworth & Yeager, 2015; Winne & Perry, 2000).

Another reactive method of assessing IER is founded on statistical analyses designed to assess the variability of an individual's responses within a survey and includes indices such as long-string analysis (Costa & McCrae, 2008) and individual response variability (IRV; Dunn et al., 2016). The underlying assumption of these indices is that, on surveys with multiple scales assessing individual difference constructs, individuals should be responding with some degree of variability. That is, when individuals respond to a great many items in a row using the same response(s) options (e.g. answering 5 to many items or answering 4,5,4,5,4), especially across scales that do not assess the same trait, they are likely not paying attention to the items or are actively looking to avoid expending cognitive effort. Curran (2016) provides an excellent review of the methods mentioned here, as well as many others that have been used to identify IER.

1.3 The Present Study

Because of the overreliance on SRS to assess and understand SRL, it is important that researchers can trust that respondents address items with the care and attention necessary to provide data for valid conclusions. Identifying instances of IER within data sets that contain items assessing the motivational and strategic aspects of SRL is one way that researchers can empirically evaluate their data and in part verify the validity of conclusions they draw regarding SRL. Further, we expand on recent work to identify IER by examining it amongst a group of college students who completed three SRS intended to assess various aspects of motivation and SRL, over the course of one academic semester. A key contribution of the present research lies in our ability examine IER in a single sample of participants, across multiple surveys measuring different constructs. From an SRL perspective, this study provides a direct examination of the possibility that students are not providing the necessary cognitive effort needed when answering items designed to assess their motivation and SRL, which threatens any conclusions using data generated by SRS. We pursued the following research questions:

(1) How prevalent is IER within a sample of college students completing SRS that assess constructs associated with SRL?
(2) Does IER manifest itself in a consistent way across different survey administrations within the same sample?

(3) Does students' engagement in IER alter the results and conclusions drawn from SRS assessing the motivational and strategic aspects of SRL?

To answer these questions, we utilized four established reactive indices of IER to: a) identify the percentage of students engaging in IER within each survey administration, as well as across survey administrations; b) examine the relationship of these indices across survey administrations; and c) identify students who engaged in IER and test whether their inclusion within the data set impacted basic univariate and multivariate statistics for motivational and strategic aspects of SRL.

2. Method

2.1 Participants

Participants were students (n = 297) at a large public university in the United States who indicated their ethnicity as White (n = 194, 64%), African-American (n = 41, 13%), Hispanic (n = 22, 7%), Asian (n = 15, 5%), and Other/Multiple (n = 29, 11%). Our sample included a majority of students who were in their first or second year of college (n = 177, 57%), had an average age of 20.4 years (SD = 4.1) and included more males (n = 168, 55%).

2.2 Procedures

Participants were recruited from 16 sections of a three-credit, letter-graded, semester-long elective course designed to improve students' SRL, and as a result, their overall academic success. As part of their assigned work for the course, students responded to four online SRS. The initial survey solicited information about how students learned about the course, their reasons for taking the course, and their knowledge about other academic outreach resources available through the university. The three remaining surveys, from which the data for the present study are drawn, were designed to assess various dispositions, beliefs, attitudes, and behaviours associated with engagement, learning, and academic success. Three-hundred and five students provided informed consent that allowed for the use of their course data for research, but eight (2.6%) of those students were missing data to an extent that precluded them from inclusion in the present study.

Procedures for each of the three relevant surveys were similar. Within the course's learning management website, students were provided a short description of the topic and purpose of the survey and concomitant assignment along with a hyperlink. When students clicked the link, a new browser window appeared displaying the first page of the particular Qualtrics-based survey. For the most part, students accessed and completed these surveys outside of the regular class period, at a time and place of their own choosing. As a final step of each survey, students were provided a "Score Report" that included a short description of their own mean scores for the relevant scales. Except for the final survey, these reports served as the basis for a personal reflection assignment and for in-class discussions, which factored into the calculation of the overall course grade. Hence, students' motivation for completing each survey was likely derived from both their interest in obtaining personal insights regarding their own motivation and strategic behaviour as well as the connection to their course grade.

2.3 Measures

The three relevant surveys (hereafter referred to as Week 2 Survey, Week 6 Survey, and Week 14 Survey for the weeks that they were assigned during the 15-week semester), each started with a short set of directions (121-171 words) and a few items (e.g. university ID numbers), that later could be used

to link them to a particular student. The initial directions identified the general topic of the survey (e.g. motivation), assured the students that there were no right or wrong answers, and provided information about how to ensure they received credit for the associated assignment. The remainder of each survey was organized into sections, each of which began with a few sentences that identified the more specific topic (e.g. self-confidence) it covered, instructed the students to read each item carefully before responding, and reminded students to be honest in their answers. For most items, students were presented with a statement and asked to indicate the extent to which it applied to themselves using a 5-point response scale ranging from *Strongly Disagree* to *Strongly Agree*. The self-efficacy for self-regulated learning items, the lone exception to this format, were answered using a 5-point response scale ranging from *Not Confident at All* to *Very Confident*.

The Week 2 Survey contained 77 items, organized into seven sections that assessed dispositions, an array of motivational beliefs, and various attitudes and behaviours related to time management and procrastination. The Week 6 Survey consisted of 43 items, organized into four sections, which assessed students' reported use of various cognitive, metacognitive, and motivational strategies. The Week 14 Survey was designed as a follow-up that would allow students to consider changes in some of their beliefs and behaviours. This final survey included 24 items from the Week 2 Survey and 18 items from the Week 6 Survey, organized into five sections. See Table 1 for a full list of the constructs assessed on each survey, as well as Appendix A for brief descriptions of these constructs and sample items.

Table 1

Description of Surveys

Construct	Source	Number of Items
	Week 2 Survey	
Mindset	Dweck (2012)	3
Intentional Delay	Choi & Moran (2010)	4
Utility Value	Hulleman et al. (2008)	4
Procrastination	Tuckman (1991)	10
Time Pressure	Created by researchers	10
Self-Efficacy for SRL	Bandura (2006)	11
Grit	Duckworth & Quinn (2009)	12
Time Management	Macan (1994)	23
	Week 6 Survey	
Environmental Management	Pintrich et al. (1993)	6
Cognitive Strategies	Pintrich et al. (1993)	12
Motivation Regulation	Wolters & Benzon (2013)	12
Metacognitive Strategies	Pintrich et al. (1993)	13
	Week 14 Survey	
Mindset	Dweck (2012)	3
Environmental Management	Pintrich et al. (1993)	6
Time Pressure	Created by researchers	10
Self-Efficacy for SRL	Bandura (2006)	11
Motivation Regulation	Wolters & Benzon (2013)	12

Note: Items on several of the scales were modified slightly in order to better reflect the particular context and/or level of specificity. In some cases, additional original items were also included. SRL = Self-Regulated Learning

2.3.1 IER Indices

Students' responses to each survey were used to compute four indices designed to assess the extent to which they engaged in IER. These indices were the primary measures used to address our research questions. Each index was computed for each survey, based on the items within that particular survey.

2.3.1.1 Response Time

Students' total response time for each survey was computed from start (when the "begin" button was clicked) to stop (when the final "submit" button was clicked). This time was automatically and surreptitiously recorded via the Qualtrics software. The total response time then, included all the time that participants used to read the directions, the time taken to read, think about, and respond to all of the items, and any time taken to read the report of their scores on the variables measured in the survey. The total response time also included any time in which the students had the survey open and active but were not actively working to complete it (e.g. were distracted).

In our effort to identify students who engaged in IER, we considered only excessively short response times. An exact minimal amount of time considered necessary for properly responding to any particular survey is difficult to determine. Huang et al. (2012) recommended a cutoff score of two seconds per item as a means of detecting IER. However, unlike surveys examined in much of the recent IER research, our surveys contained not only items and response options, but also directions and descriptions of each variable being assessed. To account for the additional time needed to read and consider all of this material, we elected to use a method of determining a cutoff score based upon expected reading rates. Carver (1982) found that the average college student can read and comprehend about 300 words per minute (wpm). Based on this rate, we computed the minimal amount of time that a student might be expected to spend completing each survey and used it to establish distinct cutoff scores for each (see Table 2).

2.3.1.2 Individual Response Variability (IRV)

The IRV index (Dunn et al., 2016) was computed by calculating a standard deviation for all items on a specific survey for each individual participant. The IRV index is sensitive to both long-string responses and less obvious forms of IER evidenced by low variability among responses (e.g. answering 4,5,4,5,4,5). Values of the IRV index could range from zero to 2.5 (the highest maximum standard deviation given our response options).

Lower values on the IRV represent less variability among a participant's responses and are interpreted as greater engagement in IER. That is, when respondents use a restricted response set like in the example above, their IRV index score will be lower and thus indicative of IER. However, no established cutoff score for IER detection exists for this index. We elected to use 2 *SD*s below the mean IRV score for all participants as our cutoff criteria. Based on the assumption of a normal distribution, this cutoff value is rather conservative, as less than three percent of scores would be expected to fall below this value.

2.3.1.3 Psychometric Synonyms (PS)

The PS index was computed as a respondent's average bivariate correlation for the set of items found to have the strongest correlations for the total sample (e.g. "I constantly worry about how little time I have for completing my assignments" and "I stress a lot about not having the time I need for my coursework") . The criteria for identifying the set of item pairs to use in the computation of the PS index is not absolute. Some have suggested identifying a particular number of item-pairs with the strongest bivariate correlations (Johnson, 2005), while others have suggested using a specific threshold for the magnitude of the correlations used in this computation (|.60|; Meade & Craig, 2012). Given the length of our surveys and the recommendation that items not be repeated when considering what item pairs to use in this computation (Curran, 2016), we computed a PS index using the 10 non-repeating pairs on

each survey with the strongest, positive bivariate correlations for the sample.[1] The PS index could range from -1 to 1 with lower scores interpreted as increased engagement in IER. However, because there is no consensus as to what value unequivocally indicates that students have engaged in IER we again elected to use 2 *SD*s below the mean value for the sample as a cutoff value.

2.3.1.4 Even-Odd Consistency (E-O)

The E-O index (Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012) is computed as a participant's correlation for the set of even and odd item-pairs that assess the same underlying construct (e.g., grit item #1 & grit item #2). The E-O index was computed using 37 item-pairs for the Week 2 Survey, 21 item-pairs for the Week 6 Survey, and 19 item-pairs for the Week 14 Survey. The range of the E-O index is also between -1 and 1, with lower scores interpreted as reflecting greater engagement in IER. Although a specific cutoff value for this index has been suggested (.30 by Jackson, 1977, as cited in Johnson, 2005), this value has not been used consistently in other IER research. Thus, to maintain consistency with the IRV and PS indices, we used 2 *SD*s below the sample mean for this index as the cutoff value to categorize students as engaging in IER.

3. Results

Table 2
Indices, Cutoff Scores, and Number of Participants Identified as Engaging in IER

	Week 2 Survey (n = 278)			Week 6 Survey (n = 281)			Week 14 Survey (n = 268)		
Index	Cutoff Value	*M* *(SD)*	ID (%)	Cutoff Value	*M* *(SD)*	ID (%)	Cutoff Value	*M* *(SD)*	ID (%)
RT (min)	5:34	240:12 (1,122:51)	5 (1.6)	4:01	110:08 (371:45)	18 (5.9)	3:18	66:02 (283:46)	22 (7.2)
IRV Index	0.66	1.05 (0.20)	8 (2.6)	0.39	0.86 (0.24)	4 (1.3)	0.46	1.00 (0.27)	4 (1.3)
PS Index	0.07	0.65 (0.28)	11 (3.6)	- 0.27	0.43 (0.34)	11 (3.6)	- 0.02	0.62 (0.31)	13 (4.3)
E-O Index	0.17	0.68 (0.27)	18 (5.9)	- 0.24	0.49 (0.53)	20 (6.6)	- 0.10	0.72 (0.41)	15 (4.9)
Total Unique ID			39 (12.8)			38 (12.5)			48 (15.7)

Note. The ID column represents the number of participants identified as engaging in IER for each method. The Total Unique ID row represents the number of individual students that engaged in IER as indicated by at least one method. It was possible (and happened to be) that a student could be identified as engaging in IER by more than one index. RT = Response Time. IRV = Individual Response Variability. PS = Psychometric Synonyms. E-O = Even-Odd Consistency

3.1 Prevalence of IER

Our first research question concerned the prevalence of IER exhibited by students who completed each of the three relevant SRS. As a first step in addressing this question, Table 2 provides the specific cutoff scores used to identify IER, the average score for each index, as well as how many students were identified as engaging in IER for each survey, based on each of the four indices utilized.

[1] A similar index based on antonyms can also be computed. However, our surveys lacked the substantial number of items with strong negative correlations necessary for this index.

Based on at least one of the indices, 12.8% of students who completed the Week 2 Survey, 12.5% of students who completed the Week 6 Survey, and 15.7% of students who completed the Week 14 Survey were identified as having engaged in IER. Overall, out of 827 survey administrations, 15.1% displayed some evidence of IER.

3.2 Consistency of IER

3.2.1 Index Consistency

Our second research question concerned the consistency of IER across survey administrations. We conceptualized consistency in two ways. First, we considered the consistency of students' values for each particular index across the three surveys. Regarding the consistency of index values across survey administrations, Table 3 displays mixed evidence. Most apparent, the values of the IRV index showed moderate to strong correlation with one another (*r*s > .43). These correlations indicate that students were somewhat consistent in whether their responses were tightly centred around a particular response option (e.g. the midpoint) or whether they tended to be more varied in the response options they selected across surveys. Students' general tendency to respond to the selected item pairs in a way that was consistent with the overall sample (i.e., the PS index) was neither exceptionally strong nor consistent.

Table 3

Correlations among IER Index Values

IER Index	1	2	3	4	5	6	7	8	9	10	11	12
					Week 2 Survey (n = 278)							
1. RT	-											
2. IRV	.04	-										
3. PS	-.02	.21**	-									
4. E-O	-.10	.19**	.27**	-								
					Week 6 Survey (n = 281)							
5. RT	**.12***	.10	-.05	.02	-							
6. IRV	.03	**.45****	-.07	-.10	.06	-						
7. PS	.00	.05	**.04**	.01	-.01	.18**	-					
8. E-O	.04	.02	.07	**.03**	.04	.12*	.30**	-				
					Week 14 Survey (n = 267)							
9. RT	**-.03**	.02	.04	.01	**.32****	.01	-.05	-.08	-			
10. IRV	-.11	**.52****	.18**	.12	.07	**.43****	.12	.02	.04	-		
11. PS	-.04	.16*	**.17****	.13*	-.02	-.02	**.16***	.10	-.10	.37**	-	
12. E-O	-.01	.11	.07	**.03**	.01	-.01	.11	**.14***	-.05	.25**	.37**	-

Note. RT = Response Time. IRV = Individual Response Variability. PS = Psychometric Synonyms. E-O = Even-Odd Consistency. Correlations in **bold** indicate the relationship between corresponding indices across surveys. * $p < .05$. ** $p < .01$.

E-O values between any of the surveys were not correlated even though, similar to the PS index, this index is a method of assessing individual response reliability. As well, evidence of consistency in the amount of time it took students to complete each survey was weak and inconsistent. That is, students who exhibited longer (or shorter) response times on one survey were no more or less likely to exhibit longer (or shorter) response times when completing the other surveys.

3.2.2 Person Consistency

As a second method of assessing consistency, we also considered whether the IER behaviour is something that students consistently engage in or is more of a one-off, situational behaviour. We examined this by looking at students who were identified as engaging in IER, by at least one index on one survey, and seeing if they were identified as engaging IER on one of the other survey administrations. Of the 297 students in our sample, just one was identified as engaging in IER on all three surveys. Further, only 27 students (9%) were identified as engaging in IER on two of the three

surveys. Of those students who were identified as engaging in IER on more than one SRS, 64% were identified by the same index on the surveys in which they were deemed inattentive. These results suggest that engagement in IER is not a consistent behaviour and, therefore, may likely be more dependent upon situational factors, such as fatigue or being distracted. When students do engage in IER repeatedly though, they tend to be inattentive in the same manner in which they had previously been inattentive.

3.3 Impact of IER on Motivation and SRL Variables

Our third and most critical research question concerned whether students' engagement in IER had an appreciable impact on basic analyses that involved the SRL and motivation variables assessed by each survey. We evaluated whether the inclusion (or exclusion) of students categorized as engaging in some form of IER in the sample altered key psychometric and descriptive properties of the motivation and SRL variables and would thus alter how this data would be interpreted. Further, we compared groups of students categorized as Attentive and Inattentive for each survey.

Table 4
Differences between the Total Sample, Attentive, and Inattentive Students on SRL Variables

	Total (n = 278, 281, 267)			Attentive (n = 238, 242, 220)			Inattentive (n = 39, 38, 48)		
Variable	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α
Week 2 Survey									
Grit [a]	3.26	0.51	.81	3.29	0.52	.82	3.10	0.43	.71
Mindset	3.52	0.94	.89	3.55	0.95	.89	3.35	0.88	.86
Utility Value [ab]	4.13	0.66	.85	4.18	0.67	.86	3.81	0.50	.65
Self-Efficacy for SRL [a]	3.77	0.53	.77	3.83	0.52	.76	3.42	0.47	.71
Time Pressure [ab]	3.24	0.78	.92	3.21	0.80	.92	3.44	0.57	.85
Intentional Delay [ab]	2.91	0.72	.70	2.84	0.72	.71	3.29	0.52	.41
Procrastination [ab]	2.89	0.66	.88	2.81	0.64	.88	3.38	0.53	.77
Time Management [a]	3.46	0.49	.86	3.50	0.48	.86	3.21	0.46	.83
Week 6 Survey									
Cognitive Strategies [b]	3.53	0.48	.72	3.52	0.46	.69	3.59	0.57	.83
Metacognitive Strategies	3.50	0.49	.80	3.50	0.48	.80	3.50	0.50	.82
Motivation Regulation [a]	3.20	0.63	.89	3.16	0.62	.89	3.42	0.63	.90
Environmental Regulation	3.65	0.67	.82	3.65	0.66	.81	3.68	0.63	.85
Week 14 Survey									
Mindset [a]	3.66	1.01	.91	3.72	0.98	.91	3.38	1.08	.92
Self-Efficacy for SRL [a]	4.09	0.56	.85	4.13	0.54	.85	3.90	0.59	.87
Time Pressure [ab]	2.99	0.84	.93	2.89	0.84	.93	3.45	0.69	.89
Motivation Regulation [b]	3.60	0.66	.93	3.59	0.68	.93	3.64	0.54	.88
Environmental Regulation	3.94	0.62	.84	3.97	0.63	.85	3.78	0.53	.78

Note. n = Week 2 Survey, Week 6 Survey, Week 14 Survey; SRL = Self-regulated learning.
[a] Significant mean level difference between Attentive and Inattentive ($p < .001$).
[b] Significant Cronbach's alpha difference between Attentive and Inattentive ($p < .05$).

3.3.1 Internal Consistency of SRL Scales

Table 4 displays Cronbach's alpha for each of the substantive measures of SRL and motivation across the three surveys for the Total sample (both Attentive and Inattentive students) as well as for Attentive and Inattentive (i.e. were identified as engaging in IER) students separately. We conducted Feldt's tests (Feldt, et al., 1987; Diedenhofen & Musch, 2016) to evaluate the observed differences in the reliability of the SRL scales computed for these three groups. Most notably, there was no statistical difference in the reliabilities computed for the Total and Attentive students. That is, the internal consistency of the SRL and motivation scales remained essentially the same, regardless of whether Inattentive students were or were not included in the computations.

In contrast, some differences in the internal consistency of these variables was observed when comparing Attentive and Inattentive students. On the Week 2 Survey, compared to Inattentive students, reliabilities for the Attentive students were significantly higher ($p < .05$) for four of the eight scales assessed (utility value, time pressure, intentional delay, and procrastination). In contrast, on the Week 6 Survey, Inattentive students had higher internal consistency for one of the four scales assessed (cognitive strategies). On the Week 14 survey, Attentive students had significantly higher consistency on two of the five scales assessed (time pressure and motivation regulation).

3.3.2 Mean response for SRL Scales

Table 4 also displays means and standard deviations for the Total sample, Attentive, and Inattentive students for each of the SRL and motivation scales on each of the three surveys. Based on independent *t*-tests, we found no statistical differences in the means for any of the SRL or motivation scales when comparing Total and Attentive students ($ps > .16$). Again, the inclusion of Inattentive students did not meaningfully change the computation of these basic descriptive statistics.

In contrast, there were several observed differences in mean scores for Attentive and Inattentive students. On the Week 2 survey, Attentive and Inattentive students showed significantly different means ($p < .001$) on utility value, self-efficacy for SRL, and time management ($M_{\text{Attentive}} > M_{\text{Inattentive}}$), as well as intentional delay and procrastination ($M_{\text{Inattentive}} > M_{\text{Attentive}}$). No significant differences in means were found for the scales on the Week 6 Survey. On the Week 14 Survey, Inattentive students reported significantly higher time pressure than did Attentive students. The observed significant differences between Attentive and Inattentive students on the SRL and motivation scales across all surveys was medium to large, based on computed effect sizes (Hedges' $g > .57$).

3.3.3 Relations among SRL Scales

Using Fisher's *r* to *z* transformation, we examined differences between the Total sample, Attentive, and Inattentive students in their patterns of correlations among the substantive SRL and motivation measures. Consistent with the findings on internal consistency and means, there were no significant differences in the correlations between SRL variables when comparing Total and Attentive students ($zs > |.65|$). Put differently, the removal of students who were identified as not providing the necessary cognitive effort needed to complete the SRS did not impact the bivariate relations between the SRL and motivation variables.

Again, we did find several differences between Attentive and Inattentive students regarding their correlations between SRL and motivation scales. On the Week 2 Survey, Attentive and Inattentive students had significantly different ($p < .05$) correlations between self-efficacy for SRL and grit, time management and procrastination, and grit and intentional delay (stronger correlations for Attentive students), as well as between intentional delay and mindset and utility value and time pressure (stronger correlations for Inattentive students). On the Week 6 Survey, Inattentive students had significantly stronger correlations between cognitive strategies and environmental management, as well as motivation regulation and metacognitive strategies. On Week 14 Survey, the only significant different in correlations between Attentive and Inattentive students was between motivation regulation and environmental management, in which Inattentive students displayed a stronger correlation.

4. Discussion

Despite their acknowledged limitations, SRS have been and remain, the most common form of assessment used to investigate SRL (Winne & Perry, 2000; Wolters & Won, 2018). Adding to this overall trend, the increasing use of online surveys continues to add to the concerns about the validity of the SRS data used to understand SRL (Karabenick & Zusho, 2011; Schellings & van Hout-Wolters, 2011; Winne & Jamieson-Noel, 2003). One major threat to validity that is consistently highlighted when considering these types of assessments is that participants do not engage in the necessary cognitive processes required to provide responses that are a valid representation of their true beliefs, attitudes, and behaviours (Duckworth & Yeager, 2015; Schwartz & Oyserman, 2001). Respondents that are inattentive or provide insufficient effort when responding to survey items are seen as corrupting influences on the resulting data that ultimately lead to invalid conclusions (Huang et al., 2015; Curran, 2016).

Our primary goal was to evaluate whether responses from inattentive students are prevalent enough to degrade the quality of data produced from SRS designed to assess key aspects of SRL and thus alter any inferences made from it. We pursued this goal by addressing three related questions about college students' engagement in IER on three SRS designed to assess multiple factors associated with SRL. In the remainder of this section, we discuss findings with regard to these questions, identify paths for future research, and make suggestions to researchers about ways they can assess the quality of data derived from online SRS.

4.1 How common is IER?

Our first objective was to evaluate the extent to which college students engaged in IER when responding to SRS designed to assess motivational and strategic aspects of SRL. Across three surveys administered over the course of a semester, we found a rate of IER (about 15%) using four common *post hoc* indices of IER that were similar to prior IER research (Meade & Craig, 2012; Maniaci & Roegge, 2014; Bowling et al., 2016; DeSimone & Harms, 2017). Much of the prior research has been conducted using personality or job-related surveys, typically on instruments with over 100 items. Our findings, therefore, support an extension of this work by demonstrating similar levels of engagement in IER when surveys are shorter and designed to assess constructs more central to the study of SRL. Hence, the theoretical content of the items may not have a strong influence on how likely respondents are to engage in IER. Perhaps even more remarkable, this consistency was found even when students were expected to have a greater personal investment in responding to items in a careful and attentive manner. Students were reminded repeatedly that the information they provided would be used for a class assignment and that it, in turn, would serve to support their positive growth and development as a student. While we assume that we were working with a more motivated sample, future research should manipulate the personal investment respondents have for responding to items and test whether this leads to differences in IER behaviours.

Our findings also exposed interesting differences based on the methods used to identify students who engaged in IER. Across survey administrations, IER was most likely to be identified based on students' score on the E-O index, followed closely by their response time. The use of the E-O consistency index, a measure of a respondent's consistency when answering items on the same scale, is primarily a check on random responding to conceptually similar items. The comparatively high proportion of students identified by this method suggests that the type of IER most common in SRL research may be more nefarious than simply speeding through the SRS or repeatedly answering with the same response option. This finding highlights the need to use more sophisticated indices to evaluate IER, rather than simply "eye-balling" SRS data for overt instances of inattentive responding.

Of the reactive methods used to identify IER, response time is arguably the most objective measure because it is founded on a less debatable premise. In particular, this method is based on the straightforward assumption that there is some minimum amount of time needed to process written text and provide a response. Based on Carver's (1982) estimation for college students, and unlike previous

studies, we used reading rate to determine an IER cutoff score for response time. It provides the advantage of considering all of the reading that accompanies taking an SRS (directions, explanations of items), factors that are unaccounted for when using Huang et al.'s (2012) recommendation of two seconds per item. Even with the advantages of using reading rate, this method of determining a cutoff score should be viewed as conservative because it does not directly consider the amount of time that a respondent would need to thoughtfully reflect upon and answer the item

Fewer students were identified as engaging in IER using the PS and IRV indices. This may be in part due to the length of the surveys we examined, as well as how the IER cutoff was determined. Previous studies that have used the IRV index identified a certain percentage of students (10%) with the lowest IRV index score to be inattentive (Dunn et al., 2016; DeSimone & Harms, 2017). Rather than predetermine the exact percent of students who must be engaged in IER, we chose instead to use a cutoff value (-2 *SD*) that is commonly used as a criterion for identifying extreme outliers in a normal distribution of scores. This criterion may underestimate who should be categorized as engaging in IER. However, as the IRV index is a relatively new method of detecting IER, more work needs to be done in order to understand its proper utilization as a method of detecting IER. In particular, evaluation of the best criteria to use for determining which students are engaged in IER would be useful.

The length of our surveys made finding items to use in calculating the PS index difficult, especially given the recommendation that items only be used once (Curran, 2016). Consequently, in order to create a reliable coefficient, we were forced to rely on only the highest, non-repeating item-pair correlations, some of which had somewhat modest correlations. The length and nature of our surveys also made it so that we could not compute a psychometric antonym index, which is typically used in conjunction with the PS index. It may be that these indices are not useful measures of IER when assessing multiple distinct constructs using scales with a relatively low number of items as is typical within the research examining SRL. Put differently, the PS index may prove more valuable for identifying IER when using scales with larger sets of items to assess cohesive underlying constructs.

Overall, these findings provide additional support for the recommendation that researchers utilize multiple indices to identify respondents who engage in IER (Huang et al., 2012; Meade & Craig, 2012). Within each survey, we found that the various indicators of IER had, at best, weak to moderate positive correlations with one another. Further, we found that very few students were identified as engaging in IER by more than one index on any particular survey. In fact, of the 125 students who were identified as being inattentive, only 14 were flagged by two or more indices within a particular survey. Overall, these findings are in line with the assumption that inattention or insufficient effort may be manifested in a variety of ways (Curran, 2016; DeSimone & Harms, 2017; Huang et al., 2012) and therefore any single method of detecting IER will fall short of identifying all the participants who engage in IER.

4.2 How consistent is IER across surveys and time?

Our second objective was to examine the consistency of students' engagement in IER across survey administrations. That is, we sought preliminary evidence of whether students' engagement in IER was more or less stable across the three surveys. As a first check on this issue, we found that the pattern of correlations between the same IER indices across different surveys was inconsistent. For example, the IRV indices across survey administrations were moderately correlated, whereas the other indices showed a much lower level of consistency. The most immediate implication of these findings is that students do not engage in certain forms of IER on a consistent basis. As well, this pattern of findings has implications for the potential causes of students' engagement in IER. If stable individual differences played a dominant role, one would expect that those students who, for example, had very quick response times on the Week 2 Survey would also display very quick response times on the Week 6 and Week 14 Surveys. In contrast, greater variability suggests that students' engagement in the various forms of IER may be the result of situational factors that are more likely to change between surveys.

Further corroboration for the importance of situational influences on inattentive survey behaviour comes from our evaluation of whether particular students or groups of students were more likely to be identified as engaging in IER. The "recidivism rate" of IER was very low; only one student was identified as engaging in IER on each of the three surveys and less than 10% of students were two-time offenders. Further, we found that no specific group of students, be those based on sex, ethnicity/race, year in school, or academic probation status, were more likely than another to be identified as engaging in IER. One more general interpretation of our findings, therefore, is that students' IER behaviour is influenced more strongly by situational features linked to a particular survey rather than by more stable demographic or individual characteristics, such as personality variables (cf. Dunn et al., 2016). Further, this conclusion suggests that the best index to use in evaluating the presence of IER should be tied to expectations about the situational factors and the type of unwanted behaviour they are likely to promote.

Our findings regarding the consistency of IER are slightly discordant with Bowling et al. (Study 1; 2016), who also studied IER consistency across different SRS administrations. Noteworthy differences in the population studied, the constructs assessed, and the nature of the repeated SRS administration make direct comparisons of these studies difficult to reconcile. Despite this, both studies suggest that IER consistency is a function of the type of survey administered, the sample answering the survey, and perhaps most importantly, situational factors (e.g. a transient environmental distraction) that influence the attention and effort participants provide when answering survey items.

4.3 Does students' engagement in IER impact conclusions about SRL?

Finally, and most critically, our findings indicate that including data from students who were deemed inattentive during the assessment process did not dramatically alter the results of some basic quantitative analyses. We compared Cronbach's alphas, means, standard deviations, and correlations of the substantive motivation and SRL measures in each survey for the Total sample (Attentive and Inattentive students) with those for only the Attentive students. Across all survey administrations, no statistically significant differences emerged. That is, the inclusion of the data from students identified as inattentive did not appear to corrupt basic analyses computed for the whole sample that are fundamental to studies of motivation and SRL. Participants who engage in IER, therefore, may add "noise" to the overall set of data (see below) but their presence does not appear to substantially corrupt the overall "signal" when considering these fundamental statistics.

Despite this overall lack of corruption, however, it would be inaccurate to say that inattentive and attentive students provided *equivalent* data. We found an array of significant mean-level, reliability, and correlational differences when comparing the Attentive and Inattentive students. A closer examination of these findings does not expose a simple or obvious pattern. In some instances, Attentive students displayed higher means, reliability, and correlation coefficients, whereas in others, this pattern was reversed. Attentive students displayed higher values for the more "desirable" motivational and SRL constructs, such as self-efficacy for SRL and time management, while also displaying lower values for less adaptive constructs such as procrastination. This pattern does not hold for all variables however, as inattentive students displayed higher mean levels of motivation regulation and higher internal consistency in their use of cognitive strategies than did Attentive students. In sum, we found clear evidence that the methods we used to evaluate IER identified some students who provided atypical response patterns that resulted in differences in some fundamental descriptive properties of the data. The way in which inattentive responding influences assessment of the underlying constructs, however, was not straightforward and needs further investigation.

In spite of the clear response set differences between Attentive and Inattentive students, our finding that the presence of data contributed by these inattentive students did not substantially degrade the quality of data collected or observed relations when assessing SRL should provide some relief to researchers. That is, our findings indicate that interpretations of past SRL research based on SRS may be relatively sound, in spite of the likelihood that there are instances of IER within the relevant data.

Further, our findings suggest that elaborate screening techniques, such as latent profile analysis (Shukla & Konold, 2018) or lengthy infrequency scales (Maniaci & Rogge, 2014) may not be necessary to use when trying to ensure the validity of students' self-reported motivation and use of self-regulated learning strategies.

4.4 Limitations

As with any exploratory research, there are several limitations to this study. The first limitation is the attrition of students over the course of the semester. The context in which our surveys were administered, as part of the coursework for a college elective course, meant that our access to students was dependent upon their continued enrollment and participation in the course. The Week 14 Survey, administered at the end of the semester, had the fewest number of participants, which is likely due to a decrease in enrollment and participation over the course of the semester. It is possible that we were unable to examine a sub-group of students who are more likely to engage in IER, those students who dropped the course or simply did not take the survey. A related limitation comes from our total sample size, just under 300 students. This modest sample size prevents strong conclusions about the generalizability of our results. Rather, our results should be viewed as an initial step to more explicitly evaluate the potential limitations of SRS data in SRL research.

The second limitation relates to the IER indices used. The *post hoc* nature of this analysis precluded us from using proactive methods of identifying IER, which have been shown to be useful in this line of research (Huang et al., 2012; Meade & Craig, 2012; Maniaci & Rogge, 2014; Huang et al., 2015). The use of response time as an indicator of IER is limited to only identifying respondents who answer too quickly. As of now, there is no agreed upon way to assess whether respondents who take too long to answer a survey are engaging in IER. Finally, our cutoff scores (300 wpm for response time and -2 *SD*s for the IRV, PS, E-O indices) have not been previously used to identify IER. We chose to utilize these cutoffs with the hope of providing a quick and simple metric to determine IER, without the use of elaborate statistical techniques, so that detecting IER may become a standard part of the data-cleaning process, such as looking for outliers or missing data. It is possible however that our simple cutoff scores reduce the complexity inherent in identifying IER.

4.5 Future Research and Implications for SRL Researchers

Our findings point to a number of additional lines of research that should be pursued in order to better understand IER and the conditions under which it may inhibit researchers' abilities to draw valid conclusions from their data. Ultimately, it would be beneficial for researchers to routinely compute and report a small set of easily understood indices, including both proactive and reactive indices, which would provide a ready metric regarding the extent to which IER is an issue within any particular set of data. Based on our measures and results, we see the E-O index and response time as easily computable, effective indices for detecting IER in SRL-related data sets. Additional research is necessary to determine the usefulness of the PS index and IRV index in these types of data sets. Beyond these indicators of IER, the implication that situational forces seem to play a more prominent role in the extent to which a respondent will engage in IER supports the need for more work to better understand the conditions that lead to more and less attentive responding by participants.

Along these lines, we also suggest that researchers should provide more detailed descriptions of the conditions under which respondents' complete SRS, such as we have done in this paper (see section *2.2 Procedures*), so that others can evaluate the extent that IER may be an issue. For example, providing information regarding specific instructions given to respondents, any incentives that respondents have to answer the SRS, and whether the SRS is completed in the presence of a researcher or without supervision, are several factors that may influence the likelihood of IER being present with collected data. In this way, typical issues such as fatigue, interest in content, distractions, along with more subtle factors such as surveillance and expected feedback can be better investigated for their impact on

participants' response behaviours. It is also worth noting that the work investigating IER needs to encompass both experimental and more "naturalistic" designs. Experimental studies in which researchers purposefully manipulate aspects of the assessment process in order to consider their effects on participants' response behaviours is essential for establishing causal connections. At the same time, studies of IER when participants complete research surveys under more typical and less controlled conditions (e.g., our sample) also are necessary for more ecologically valid conclusions.

5. Conclusion

Although they remain a popular choice for researchers, there are a number of important limitations that threaten the validity of using SRS to investigate SRL (Karabenick & Zusho, 2011; Schellings & Van Hout-Wolters, 2011; Winne & Jamieson-Noel, 2003). In the present study, we focused on evaluating just one of these critiques, that students' inattention or insufficient effort while completing the items on SRS will substantially reduce the integrity of the resulting data and, therefore, its usefulness for investigating SRL. Our findings lead to two overall insights regarding this critique. On the one hand, we found evidence that a notable proportion of students engaged in IER and, as a result, produced data with some basic statistical properties that were inconsistent with those produced by students who appeared to complete the surveys more thoughtfully. On the other hand, we also found evidence that the irregular response patterns or "noise" contributed by the students who engaged in IER did not corrupt the data to an extent that basic "signals" or statistical properties were lost or debased. In sum, researchers examining SRL should likely consider IER more as a nuisance that should be reduced whenever and in as many ways as possible, rather than as fatal flaw that precludes the use of SRS as a viable methodology.

Keypoints

- The prevalence of IER across three surveys assessing aspects of SRL was about 15%
- Attentive and Inattentive students provided data that was significantly different from one another
- Data from Inattentive students did not degrade descriptive statistics computed for whole sample
- Separate detection methods identified different students as engaging in IER
- Participants' engagement in IER was a function of situational influences more than individual differences

References

Azevedo, R., Taub, M., & Mudrick, N.V. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. Schunk & Greene, J.A (Eds.), *Handbook of self-regulation of learning and performance (2nd ed., pp. 254-270).* New York, NY: Routledge.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In *Self-efficacy beliefs of adolescents* (pp. 307–337). https://doi.org/10.1017/CBO9781107415324.004

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied*, *123*(1), 101–103. https://doi.org/10.1080/00223980.1989.10542966

Biswas, G., Baker, R,, & Paquette, L. (2018). Data mining methods for assessing self-regulated learning. In D. H. Schunk & J. A. Greeene (Eds.), *Handbook of self-regulation of learning and performance (2nd ed., pp. 388 - 403)*. New York: Routledge.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218-229. https://doi.org/10.1037/pspp0000085

Butler, D. L. (2002). Qualitative approaches to self regulated learning: Contributions and challenges. *Educational Psychologist*, *37*, 59–63. https://doi.org/10.1207/S15326985EP3701

Carver, R.P. (1992). Reading rate: Theory, practice, and practical mplications. *Journal of Reading, 36*(2), 84-95.

Choi, J. N., & Moran, S. V. (2010). Why not procrastinate? Development and validation of a new active procrastination scale why not procrastinate. *The Journal of Social Psychology*, *149*, 37–41. https://doi.org/10.3200/SOCP.149.2.195-212

Cleary, T. J, & Callan, G. (2018). Assessing self-regulated learning using microanalytic methods. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed.). New York: Routledge.

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools*, *41*(5), 537–550. https://doi.org/10.1002/pits.10177

Costa, P. T., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 - Personality Measurement and Testing* (pp. 179–198). London: SAGE Publications Ltd. https://doi.org/10.4135/9781849200479.n9

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

DeSimone, J. A., & Harms, P. D. (2017). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, 1–19. https://doi.org/10.1007/s10869-017-9514-9

Diedenhofen, B., & Musch, J. (2016). cocron : A web interface and R package for the statistical comparison of Cronbach ' s alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60. http://www.ijis.net/ijis11_1/ijis11_1_diedenhofen_and_musch.pdf

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, *3*(3), 231–264. https://doi.org/10.1007/s11409-008-9029-x

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (Grit – S). *Journal of Personality Assessment*, *3891*, 166–174. https://doi.org/10.1080/00223890802634290

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2016). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, 1–17. https://doi.org/10.1007/s10869-016-9479-0

Dweck, C. (2012). *Mindset: How You Can Fulfil Your Potential*. London: Robinson.

Feldt, L. S., Woodruff, D. J., & Salih, F. a. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, *100*(2), 398–416. https://doi.org/10.1037/0022-0663.100.2.398

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*, 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Karabenick, S. A., & Zusho, A. (2011). Examining approaches to research on self-regulated learning: conceptual and methodological considerations. *Metacognition and Learning*, *10*(1), 151–163. https://doi.org/10.1007/s11409-015-9137-3

Macan, T. H. (1994). Time management: Test of a process model. *Journal of Applied Psychology*, *79*(3), 381–391. https://doi.org/10.1037/0021-9010.79.3.381

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*(1), 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

McCardle, L., & Hadwin, A. F. (2015). Using multiple, contextualized data sources to measure learners' perceptions of their self-regulated learning. *Metacognition and Learning*, *10*(1), 43–75. https://doi.org/10.1007/s11409-014-9132-0

McKibben, W. B., & Silvia, P. J. (2017). Evaluating the distorting effects of inattentive responding and social desirability on self-report scales in creativity and the arts. *Journal of Creative Behavior*, *51*(1), 57–69. https://doi.org/10.1002/jocb.86

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 239–250. https://doi.org/10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*(APR), 1–28. https://doi.org/10.3389/fpsyg.2017.00422

Pajares, F. (2007). Motivational role of self-efficacy beliefs in self-regulated learning. In B. J. Zimmerman, & D. H. Schunk (Eds*.), Motivation and Self-Regulated Learning: Theory, Research, and Applications* (pp. 111-140). New York: Erlbaum.

Perry, N. E., Hutchinson, L. R., Yee, N., & Maatta, E. (2018). Advances in understanding young children's self-regulation of learning. In D. H. Schunk & J. A. Greeene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed.). New York: Routledge.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning suestionnaire (MSLQ). *Educational and Psychological Measurement*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Pintrich, P. R., & Zusho, A. (2007). Students' motivation and self-regulated learning in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 731–810). New York: Springer.

Schellings, G., & van Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments:

Theoretical and empirical considerations. *Metacognition and Learning*, *6*(2), 83–90. https://doi.org/10.1007/s11409-011-9081-9

Schunk, D. H., & Zimmerman, B. J. (1998). *Self-regulated learning: From teaching to self-reflective practice*. *Psychological Science*. Guilford Press.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behaviour. *American Journal of Evaluation*, *22*(2), 127–160. https://doi.org/10.1177/109821400102200202

Shukla, K., & Konold, T. (2018). A two-step latent profile method for identifying invalid respondents in self-reported survey data. *Journal of Experimental Education*, pp. 1–16. https://doi.org/10.1080/00220973.2017.1315713

Tuckman, B. W. (1991). The development and concurrent validity of the procrastination scale. *Educational and Psychological Measurement*, *51*(2), 473–480. https://doi.org/10.1177/0013164491512022

Usher, E., & Schunk, D. H. (2018). Social cognitive thoretical perspective of self-regulation. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed.). New York: Routledge.

Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Mahwah, NJ: Erlbaum Associates.

Winne, P. H., & Jamieson-Noel, D. (2003). Self-regulating studying by objectives for learning: Students' reports compared to a model. *Contemporary Educational Psychology*, *28*(3), 259–276. https://doi.org/10.1016/S0361-476X(02)00041-3

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In *Handbook of Self-Regulation* (pp. 531–566). Elsevier. https://doi.org/10.1016/B978-012109890-2/50045-7

Wolters, C. A., & Benzon, M. B. (2013). Assessing and predicting college students use of strategies for the self-regulation of motivation. *Journal of Experimental Education*, *81*(2), 199–221. https://doi.org/10.1080/00220973.2012.699901

Wolters, C. A., & Won, S. (2018). Validity and the use of self-report questionnaires to assess self-regulated learning. In D. H. Schunk & J. A. Greeene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed.). New York: Routledge.

Zimmerman, B. J. (2008). Investigating self-Regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, *45*(1), 166–183. https://doi.org/10.3102/0002831207312909

Appendix A

Construct	# of Items	Description of Construct	Example Items
		Week 2 Survey	
Mindset (Dweck, 2012)[a]	3	Belief that intelligence is something that can be changed	"You can learn new things, but you can't really change your basic intelligence" "Your intelligence is something about you that you can't change very much"
Intentional Delay (Choi & Moran, 2010)	4	Preference for delaying work in order to increase productivity	"To use time more efficiently, I deliberately postpone some school tasks" "I finish most of my assignments right before the deadlines because I choose to do so"
Utility Value (Hulleman et al., 2008)	4	Beliefs about the relevance of coursework for future goals	"I can apply what we are learning in my classes to real life" "I think what we are studying in my courses this term is useful for me to know"
Procrastination (Tuckman, 1991)	10	Needless delaying of work	"I postpone getting started on things I don't like to do" "I manage to find an excuse for not doing my schoolwork"
Time Pressure (created by researchers)[a]	10	Feelings of lack of control over time in daily life	"I often feel anxious about not having enough time for schoolwork" "I often wish there was more time in my day for schoolwork"
Self-Efficacy for SRL (Bandura, 2006)[a]	11	Confidence in the ability to self-regulate one's learning	"I am confident that I can participate in class discussions" "I am confident that I can finish my homework assignment by deadlines"
Grit (Duckworth & Quinn, 2009)	12	Perseverance and pursuit of long-term goals	"I am a hard worker" "I often set a goal but later choose to pursue a different one"
Time Management (Macan, 1994)	23	Beliefs and behaviours about the effective use of time in goal-directed work	"I make a list of things to do each day "I block out time in my day to work on class assignments"

Week 6 Survey

Environmental Management (Pintrich et al., 1993)[a]	6	Manipulation of context to increase productivity	"I usually study in a place where I can concentrate on course work" "When I study, I try to get rid of any distractions that are around me"
Cognitive Strategies (Pintrich et al., 1993)	12	Use of rehearsal, elaboration, and organization while studying	"I make figures, charts, or tables that help me study course materials" "I link whatever I am studying to something relevant in my life"
Motivation Regulation (Wolters & Benzon, 2013)[a]	12	Use of strategies to increase motivation while studying	"Even when studying is hard, I can figure out a way to keep myself going" "If I lose interest in an assignment, I have ways to boost my effort to get it done"
Metacognitive Strategies (Pintrich et al., 1993)	13	Use of planning, regulating, and monitoring while studying	"When I am reading I stop once in a while and reflect on what I am learning" "If an assignment is giving me trouble, I change the way I get it done"

[a] Items were repeated to create Week 14

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 126 - 139
ISSN 2295-3159

Experience and Meaning in Small-Group Contexts: Fusing Observational and Self-Report Data to Capture Self and Other Dynamics

Christine Calderon Vriesema[a], Mary McCaslin[b]

[a] University of California, Santa Barbara, USA

[b] University of Arizona, USA

Article received 17 May 2019 / Article revised 15 November / Accepted 1 January 2020 / Available online 30 March

Abstract

Self-report data have contributed to a rich understanding of learning and motivation; yet, self-report measures present challenges to researchers studying students' experiences in small-group contexts. Rather than using self-report data alone, we argue that fusing self-report and observational data can yield a broader understanding of students' small-group dynamics. We provide evidence for this assertion by presenting mixed-methods findings in three sections: (a) self-report data alone, (b) observational data alone, and (c) the fusion of both data sources. We rely on 101 students' self-reported experiences as well as observational (i.e., audio) data of students working in their group (N = 24 groups). In section order, we found that (1) students' self-reported small-group behavior predicted their end-of-study reported anxiety and emotion; (2) coded observational data captured five types of group dynamics that students can engage in; and (3) students' initial group-level characteristics predicted their real-time group dynamics, and observed group regulation activity predicted students' self-reported anxiety, emotion, and regulation moving forward. Thus, while self-report and observational data alone can each increase our understanding of student motivation and learning processes, pursuing both in tandem more effectively captures the give-and-take among students, how these experiences evolve over time, and the personal meanings they can afford.

Keywords: Self-Report; Observation; Small-Group Dynamics; Motivation; Co-Regulation

Corresponding author: Christine Calderon Vriesema is now at the University of Wisconsin-Eau Claire, Psychology Department, Hibbard Humanities Hall 260, 124 Garfield Ave., Eau Claire, WI 54701, USA. Email: vriesecn@uwec.edu DOI: https://doi.org/10.14786/flr.v8i3.493

1. Introduction

Instruments measuring students' motivation and learning processes have contributed to a rich understanding of students' experiences in school. Yet, the extent to which self-report measures adequately capture the learning process for all students across varying contexts remains an important concern (e.g., Urdan & Bruchmann, 2018). For researchers studying specific instructional contexts, self-report data can pose challenges to investigating motivation and strategy use in small groups. Namely, self-report measures make it difficult to investigate how students' reported behavior and emotion occur in real-time and in relation to other people in their immediate environments. When students work together, each person brings unique experiences and characteristics into their small groups. How identity, disposition, motivation, and readiness to learn impact group functioning—and how individuals are impacted by their interactions with others over time—reflects a dynamic process that self-report data alone cannot capture.

To better understand this complex learning environment, we pursued a longitudinal, mixed-methods study of 101 students' small-group experiences during six math lessons (n = 24 groups from two third-grade and two fifth-grade classrooms). Students completed self-report measures at pretest and at posttest. Throughout the study, students also completed an instrument describing their individual small-group behaviors after each lesson. Finally, after completing the study, students responded to items asking them how they would feel if their teacher asked them to get into small groups again. In addition to these self-report data, our project included real-time audio data of students working in their small groups.

Selected results of this study were briefly discussed as part of a larger chapter focusing on the guiding theoretical perspective (McCaslin & Vriesema, 2018); we present the full study here for the first time. The present special issue aims to better understand the impact of self-report data on theory and practice (Fryer & Dinsmore, 2020). We contribute to this goal by specifically addressing two of the three guiding questions: How does the use of self-report constrain the analytical choices made with self-report data, and how do the interpretations of self-report data influence interpretations of findings? We situate both questions within the context of small-group research. We begin by briefly introducing the guiding theory. We then present our study's findings in three sections depicting what we learn from self-report data alone, observational data alone, and integrating both data sources.

1.1 Theory

The co-regulation model (McCaslin, 2009) that guides this research is a motivation perspective positing that learners are social, have a basic need for participation and validation (McCaslin & Burross, 2008), and differ in how and what they participate (McCaslin et al., 2016). Influenced by Vygotskian tenets, this theory describes how three sources of influence function together to inform emergent identity. These sources are cultural (e.g., norms, challenges), social (e.g., relationships, opportunities), and personal (e.g., readiness to learn, disposition). Students bring their personal backgrounds and characteristic adaptations to the classroom; yet, the opportunities presented to students and the relationships formed throughout their schooling experiences can shape who students become.

Given the dynamic processes described within this theoretical perspective, small-learning groups present an opportune setting to study emergent identity. Students in small groups each bring varying achievement levels, dispositions, and motivation to the task. However, the nature of the small-group instructional setting requires that students work together toward a common goal and negotiate challenges when necessary. When students work with each other across multiple occasions, small groups provide an opportunity to understand how student identity informs their work with other classmates and how these shared classroom experiences can shape student identity moving forward.

Some scholars have hailed small-group learning formats as the success story of educational psychology (Johnson & Johnson, 2009). Small-group activities can enhance student thinking and learning of both formal (e.g., math) and informal (e.g., appropriate social skills, motivated student engagement) content and skills (e.g., Elias & Schwab, 2006; Hadwin, Järvelä, & Miller, 2018; Webb, 2008). However, while small-group learning has demonstrated benefits, there also are concerns that not all small-group activities are beneficial nor do all group members experience them similarly (Rogat, et al., 2013; Webb, 2013). Naturalistic observational studies examining the processes that actually occur within small groups and what students make

of them are relatively scarce. Extant research, however, suggests their importance (e.g., Hadwin & Järvelä, 2011; Tan et al., 2005; Webb, 2013).

Therefore, the dynamic processes occurring within small-group settings necessitate dynamic methodologies to study them. Asking students about their experiences in small groups can yield important information regarding students' interpretations of events; and, researchers can investigate how students' personal characteristics associate with these self-report data. However, self-report data alone cannot capture the give-and-take of small-group interactions. Yet, observational data alone also can fail to capture the full student experience. In the case of observation-only data, researchers rely on their own interpretations of events and fail to capture students' own self-reported experiences of the events. Thus, combining self-report and observational data provides a foundation for more fully understanding how individual characteristics inform small-group dynamics, and how these dynamics inform student identity moving forward. To illustrate these points in finer detail, we present three sections that discuss (a) self-report data, (b) observational data, and (c) the fusion of both data sources in our research.

2. Section 1: Self-Report Data

This section relies on self-report data to illustrate how students' reported small-group behavior associated with their characteristics at pretest and posttest. First, we describe how students' pretest characteristics—their teacher-ranked math readiness, self-reported anxiety and emotional adaptation (i.e., context-dependent emotion and coping strategies; McCaslin et al., 2016)—associated with their self-reported small-group behavior. Second, we show how self-reported group behavior predicted students' posttest anxiety, emotional adaptation, and reports of how they would feel if their teachers asked them to get into small groups again. To contextualize these results, we first describe the relevant method information.

2.1 Procedure

Students ($N = 101$) completed the pretest (October) and posttest (January) surveys that measured their anxiety and emotional adaptation. Teachers also ranked each of their students on mathematics achievement at pretest. At the end of the study, students completed an instrument asking them how they would feel if their teacher asked them to get into groups again. Throughout the study, students also completed short instruments immediately after each small group lesson to indicate their behavior during the lesson. We present students' average reported behavior (i.e., the average across the six lessons) below in order to enhance clarity of the results.

2.2 Data Sources

2.2.1 What School is Like for Me (WSLM)

WSLM is the *Test Anxiety Scale*, a well-known, well-researched, and well-critiqued instrument (Pekrun, 2006; Zeidner & Matthews, 2005) adapted from Sarason, et al., (1958). WSLM asks students to agree or disagree with 18 sentences describing anxious thoughts and feelings. Cronbach's alpha for WSLM was $\alpha = .76$ at pretest and .71 at posttest.

2.2.2 School Situations (SS)

School Situations (SS; Burggraf, 1993) is an adaptation of the *Test for Self-Conscious Affect* (TOSCA), a dispositional measure originally designed for adults and subsequently revised by Tangney and colleagues to include children (e.g., Tangney et al., 1995). The SS inventory asks students to use a five-point scale to endorse sentences in response to 12 written vignettes that portray routine school challenges within three contexts: whole class, small group, or private/individual. Sentences are behavioral representations of emotions (guilt, shame, or pride) and coping strategies (externalize, normalize).

Rather than consider the five SS scales (pride, guilt, shame, normalize, externalize) independently, as originally designed, we used five unique emotional adaptation profiles identified in previous research

(McCaslin, et al., 2016) for our analyses. The five profiles were: (1) *Distance and Displace*: the student attempts to withdraw from a difficult situation to care for the self and/or attempts to blame other people or things to find relief from feelings of shame; (2) *Regret and Repair*: the student attempts to repair or fix the situation and to care for the self through normalizing the event in order to find relief from feelings of guilt; (3) *Inadequate and Exposed*: the student assumes responsibility and blame for mistakes or difficulties without engaging in self-care or displacement strategies in response to negative emotion; (4) *Proud and Modest*: the student acknowledges success, but tempers feelings of pride with humility; and (5) *Minimize and Move On*: the student adopts a 'just keep going, do not dwell, look beyond it' escape response to mistakes and difficult situations. At pretest, Cronbach's alpha was .75, .79, .70, .72, and .64 for Distance and Displace, Regret and Repair, Inadequate and Exposed, Proud and Modest, and Minimize and Move On, respectively. In the same order, internal consistency reliability at posttest was .75, .87, .70, .75, and .66, respectively.

2.2.3 How I was in Group Today (How I Was)

How I Was presented 20 sentences to students and asked them to underline any that described their behavior in their group that day. Sentences comprised three scales (McCaslin, et al., 1994): (1) *Enhancing:* sentences that represent engagement from which other group members may benefit; (2) *Neutral*: sentences that represent participation that is neither active nor withdrawn; and (3) *Interfering*: sentences that describe preoccupation with concerns of the self. The Interfering scale consisted of items suggesting that students withdrew from or were unable to participate in small group activity (e.g., "My stomach felt funny"; "My head hurt") rather than engaging in behaviors that actively distracted or interfered with others in small group. Therefore, we subsequently refer to this scale as "*Withdrawn"* to clarify this distinction.

2.2.4 How I Felt

How I Felt was designed to capture students' thoughts and feelings when the teacher said it was time to get into their small group. It consisted of six items that described positive and negative emotional experiences in three relative domains: cognitive, affective, and physiological. Interested (cognitive), Happy (affective mood), and Relaxed (physiological) comprised the "positive" scale ($\alpha = .85$). Confused (cognitive), Sad (affective mood), and Nervous (physiological) comprised the "negative" scale ($\alpha = .82$). Students used a 3-point scale (not at all, a little bit, a lot) to respond to each item.

2.3 Results

2.3.1 Pretest Student Characteristics and Self-Reported Small-Group Behavior

Students' pretest anxiety and emotional adaptation did not associate with students' self-reported small-group behavior. However, students with higher initial math readiness reported greater use of Neutral regulation strategies, such as listening, during their small groups ($r = -.27$, $p = .007$; higher numbers indicate lower rank in math readiness).

2.3.2 Self-Reported Small-Group Behavior and Posttest Student Characteristics

We pursued a series of multiple regression analyses that controlled for students' reported pretest anxiety and pretest emotional adaptation. We did not control for group membership (e.g., using fixed effects models) because we believed that this might yield decontextualized results. In this paper, we focused on exploring how group processes shaped individual processes and vice versa; thus, we did not account for group membership in order to work toward this goal. However, we did attempt to cluster errors at the group level in our regression analyses in order to account for the shared variance within groups. Unfortunately, we did not have a sufficient number of participants for the number of groups in our study to run this analysis effectively. As a result, we proceeded to use traditional multiple regression analyses here and subsequently in the paper.

Results indicated that students' self-reported behavior in their small-groups predicted students' posttest anxiety ($F(9, 72) = 4.16$, $p < .001$; $r^2 = .34$, adjusted $r^2 = .26$), as well as two emotional adaptation profiles: Regret and Repair ($F(9, 71) = 5.22$, $p < .001$; $r^2 = .40$, adjusted $r^2 = .32$) and Inadequate and Exposed ($F(9, 71) = 3.00$, $p = .004$, $r^2 = .28$, adjusted $r^2 = .18$). Specifically, reported use of Enhancing regulation during small group predicted less anxiety at posttest ($\beta = -2.32$, $p = .035$). Use of Withdrawn regulation also predicted

lower endorsement of Regret and Repair and Inadequate and Exposed emotional adaptation at posttest (β = -0.25, p = .01; β = -0.22, p = .051, respectively).

2.3.3 Self-Reported Small-Group Behavior and Posttest Anticipated Affect

We pursued a series of multiple regression analyses that controlled for students' pretest anxiety and emotional adaptation to determine how students' self-reported behavior during small group predicted their anticipated affect at posttest (i.e., when they imagined the teacher asking them to get into small groups again). Students' self-reported behavior in small groups predicted their endorsement of both positive and negative affect ($F(9,75) = 3.00$, $p < .001$, $r^2 = .32$, adjusted $r^2 = .24$; $F(9,75) = 3.45$, $p = .001$, $r^2 = .29$, adjusted $r^2 = .21$, respectively). Reported Enhancing behavior during small group predicted greater anticipated positive affect (β = 0.50, $p < .001$); in contrast, reported Withdrawn behavior predicted greater anticipated negative affect (β = 0.42, $p < .001$).

2.4 Constrained Analytical Choices and Interpretations

Overall, the self-report data indicated how students' reported small-group behavior associated with their personal characteristics and attitudes at pretest and posttest. Specifically, average student-perceived Enhancing behavior across the six lessons predicted lower anxiety at posttest and greater positive affect at the end of the study when students imagined getting into small groups again. In contrast, students who described themselves as Withdrawn during their small groups felt more negative emotion when they imagined getting into small groups again. Student-perceived Withdrawn behavior also predicted less endorsement of Inadequate and Exposed and Regret and Repair emotional adaptation; thus, while withdrawing from participation might mitigate the potential for experiencing shame in small-group settings, it also prevents students from potentially developing strategies for overcoming interpersonal challenges with peers.

Although the interpretations of self-report data provided insight into how students' perceived small-group behavior associated with their personal characteristics and expectations (e.g., affect), there are several important limitations. First, our analyses were constrained by individual-level data. The data allowed us to examine how students' self-reported behavior associated with their pretest and posttest outcomes; yet, students do not participate in their small groups alone. The constrained data sources prevented a more complete understanding of the give-and-take *among* students in these settings. Second, our interpretations of the data relied purely on student reports. Students' individual interpretations of their classroom activities are vital to understanding their emergent identity; however, finding ways to corroborate self-report data with real-time data can enhance understanding of self- and other-awareness in small-group dynamics.

3. Section 2: Observational Data

While Section 1 illustrated associations with students' *self-reported* small-group behavior, Section 2 depicts students' *actual* behavior during their small groups. In Section 2, we describe the types of co-regulation dynamics that emerged during students' small-groups lesson and how the dynamics associated with the groups' average achievement on the small-group tasks; the group is the unit of analysis. We present the observational results after describing the relevant procedures and coding systems.

3.1 Procedure

Three researchers independently analyzed, transcribed, and verified audio data of small-group interactions for three lessons (representing the beginning, middle, and end of the six lessons) for each group (N = 24 groups). The three researchers remained unaware of the larger study. Two complementary observation systems were developed for analyzing the audio data**.** We describe the coding systems below.

3.2 Data Sources

3.2.1 Group Behavior Checklist (GBC)

The first system, the GBC, is a lower-inference observation instrument that captured the range of on- and off-task behaviors that students displayed when working with others in small groups. This study used four GBC variable domains: (a) planning, (b) problem solving, (c) help-seeking, and (d) feedback. Coding was completed in 30-second intervals. In total, 2,180 intervals were coded with the GBC. The percentage of exact agreement (91%) among coders was calculated on three coding pairs over three lessons.

3.2.2 Group Environment Summary (GES)

The second system, the GES, is a higher-inference system that captured students' interpersonal and affective dynamics and expressed intrapersonal coping strategies. Variable domains included group affective climate; giggle/laugh bursts; and types of aggressive, protective, regressive/escape, and somatic expressed coping behaviors. Coding was completed in two halves: at the mid-point and end of each lesson. The percentage of exact agreement was 73% among three coding pairs over three lessons. See McCaslin and colleagues (2011) for more complete documentation of audio enhancement and transcription procedures; McCaslin and Vega (2013) for coding system design, procedures, and application in the pilot study; and Vega (2014) for implementation decisions for the revised system.

3.2.3 Group Achievement

Student activity worksheets completed "by the group" as part of each lesson were scored and verified by two math educators for correctness. Percentage correct represented students' small-group achievement for the lesson material.

3.3 Results

3.3.1 Group dynamics

We represented the GBC and GES data as the percentage of intervals in which a behavior occurred. We then subjected the data from both observation systems to a principal components analysis using Varimax rotation in order to develop an understanding of overall group dynamics from our discrete coding categories. Results yielded five independent factors that collectively accounted for 61.92% of the variance in student small-group interactive behavior. Factors, in order of magnitude, were: Conflict and Control, Working Together, Resource Drain, Edgy Compliance, and Scuffle and Confusion (see Table 1 for example behaviors and percentage variance explained by factor). We consider these five distinct co-regulation dynamics that students can engage in while in small groups.

Table 2
Real-Time Co-Regulation Dynamics

Factor (% Variance)	Item	Loading
1. Conflict and Control (25.96%)	Students are indifferent/non-cohesive/within group coalitions exist	.84
	Students display unidirectional aggression – off-task	.84
	Students display reciprocal aggression – off-task	.77
2. Working Together (11.62%)	Students offer explanations	.67
	Students express disagreement	.66
	Students ask procedural questions	.64
3. Resource Drain (9.34%)	Students make excuses	.81
	Students make request for an audience	.66
	Students make request for materials	.66
4. Edgy Compliance (7.81%)	Students giggle/laugh	.74
	Students brag	.67
	Students refuse others' further participation/contributions	.62
5. Scuffle and Confusion (7.20%)	The source of task participation structure is not identifiable	.93
	Task participation structure is not identifiable	.74
	Students are argumentative	.66

Note. Only the top three positively loaded items for each factor are listed in the table. Total number of items varied by factor: $n = 18$ for Factor 1; $n = 11$ for Factor 2; $n = 7$ for Factor 3; $n = 9$ for Factor 4; $n = 5$ for Factor 5. The exploratory factor analysis yielded 8 cross-loaded items: 5 items loaded in the opposite direction, and 3 items loaded in the same direction.

The five small-group dynamics factors can be organized into relatively task-focused, other-focused, or the fusion of the two perspectives. In task-focused contexts, student dynamics primarily centered on the academic activity at hand, whereas dynamics in other-focused contexts reflected an emphasis on one's group members. In addition, we can consider how types of coping behaviors typically associated with individual

student behavior—aggressive, protective, and regressive—emerged as characteristics of group co-regulation dynamics. Please see Figure 1 for a visual representation of how the joint activity varied across the five small-group dynamics.

		Relative Focus of Joint Activity	
		Task-Involved	**Other-Involved**
Social/Emotional Context	**Protective**	Working Together	
	Aggressive	Edgy Compliance	Conflict and Control
	Regressive	Scuffle and Confusion	Resource Drain

Figure 1. Small-Group Regulation Foci
Note. This figure was adapted from McCaslin and Vriesema (2018).

The *Working Together* dynamic fused the demands of task and peers in small group learning within a protective press. Group members could ask for assistance, disagree with each other, and offer suggestions and solutions without concern for personal safety. In comparison, an aggressive press encompassed both the relatively task-involved *Edgy Compliance* dynamic (in which provocative and aggressive behaviors were related to attempts to meet task demands) and the other-involved *Conflict and Control* dynamic (in which aggression and protection behaviors consumed group attention). Finally, the *Scuffle and Confusion* dynamic in the disorganized pursuit of task demands and the *Resource Drain* of needy peers were each marked by regressive, or relatively immature, co-regulation dynamics. Taken together, these profiles did not represent particular groups per se; rather, they represented the types of co-regulation dynamics—that is, the types of observed behavior (e.g., communication patterns, coping strategies)—that emerged during the students' time in small groups. Please see Table 2 for the means and standard deviations for the co-regulation dynamics.

Table 3

Descriptive statistics for students' co-regulation dynamics

Variable	***M* (*SD*)**
Conflict and Control	34.77 (14.01)
Working Together	29.58 (7.44)
Resource Drain	19.93 (9.56)
Edgy Compliance	26.60 (9.76)
Scuffle and Confusion	40.88 (7.32)

Note. Means and standard deviations reflect the percentage of time students spent engaging in each of the co-regulation dynamics.

3.3.2 Group achievement

Scuffle and Confusion negatively associated with the average percentage correct on group task activities (r = -.46, p = .024). Group achievement did not associate with Conflict and Control (r = -.02, p = .929), Working Together (r = .09, p = .670), Resource Drain (r = .08, p = .718), or Edgy Compliance (r = .21, p = .321).

3.4 Constrained Analytical Choices and Interpretations

We presented this section depicting observational data alone for two reasons. First, the extensive coding systems provided a framework for understanding the types of dynamics that can emerge in small groups. The researchers observed student behavior that informed behavioral co-regulation ranging from the 'ideal' Working Together dynamics to the aggressive give-and-take between students (e.g., Conflict and Control) to the disorganized task pursuits that embodied Scuffle and Confusion. These interpretations, therefore, yielded a broader understanding of students' systematically observed behavior during their small-group lessons.

Second, we presented the observational data alone to illustrate that even with real-time data of students working in their small groups, we fail to understand what these dynamics can mean for students' identity moving forward. Relying on observational data alone constrained our analyses to correlations between small-group dynamics and group achievement on the small-group tasks. While this has the important benefit of aligning with the types of data available to teachers when they use small groups in their instruction, we argue that fusing observational and self-report data can provide a more nuanced understanding of students' experiences in small groups as well as insight into what these classroom experiences can mean for students' emergent identity. We provide evidence for this argument in the next section.

4. Section 3: Fusing Self-Report and Observational Data

Rather than constraining self-report data to individual-level analyses or relying solely on group-level observational data, Section 3 first illustrates how group-level characteristics associated with real-time group dynamics. Specifically, we took the average of group members' individual characteristics, such as emotional adaptation, to determine how the group's overall approach to learning tasks associated with their group functioning during the small-group lessons. Second, we describe how the give-and-take of small-group dynamics predicted students' individual characteristics at posttest (i.e., their anxiety, emotional adaptation, and anticipated affect). Third, while we have confidence in the reliability of our systematic coding procedures, researcher perceptions of small group dynamics may not coincide with student perceptions of their small-group experiences. Therefore, we also present results that depict the alignment between student- and researcher-perceived group behaviors.

4.1 Results

4.1.1 Pretest Group Characteristics and Real-Time Group Co-Regulation Dynamics

We created group-averaged scores for anxiety, emotional adaptation, and math readiness in order to determine how group composition associated with co-regulation dynamics.

Anxiety. Group-averaged anxiety did not associate with students' group regulation dynamics.

Emotional Adaptation. Group-averaged endorsement of Inadequate and Exposed positively associated with Working Together dynamics (r = .47, p = .021). Group-averaged endorsement of Distance and Displace positively associated with engagement in Resource Drain dynamics (r = .41, p = .048). Finally, group-averaged endorsement of Proud and Modest negatively associated with Edgy Compliance group dynamics (r = -.42, p = .039).

Math Readiness. Groups with a greater concentration of higher-ranked math students were more likely to display Working Together and Edgy Compliance dynamics (r = -.21, p = .049; r = -.23, p = .028 respectively). In contrast, groups with a greater concentration of lower-ranked math students were more likely to engage in Scuffle and Confusion dynamics (r = .37, p = .001).

4.1.2 Real-Time Group Co-Regulation Dynamics and Posttest Student Outcomes

We pursued a series of multiple regression analyses controlling for students' pretest characteristics to determine how real-time group dynamics predicted students' self-reported anxiety, emotional adaptation, and anticipated affect at posttest.

Anxiety. Small-group dynamics predicted students' self-reported anxiety, $F(11, 66) = 3.97$, $p < .001$; $r^2 = .40$, adjusted $r^2 = .30$. Specifically, participating in groups that displayed Working Together and Resource Drain co-regulation dynamics predicted lower posttest student anxiety ($\beta = -.27$, $p = .011$; $\beta = -.27$, $p = .012$, respectively). Although students may have used different strategies in the two co-regulation dynamics, receiving help from peers in both contexts may have associated with lower anxiety at posttest.

Emotional Adaptation. Group dynamics predicted students' endorsement of Regret and Repair at posttest, $F(11, 65) = 4.09$, $p < .001$; $r^2 = .41$, adjusted $r^2 = .31$. In particular, Edgy Compliance dynamics predicted greater Regret and Repair at posttest ($\beta = .21$, $p = .038$). Group dynamics did not predict the other four emotional adaptation profiles.

Anticipated Affect. We explored whether small-group dynamics predicted how students would feel if their teachers asked them to get into their small groups again. Students' observed small-group dynamics predicted their self-reported anticipated positive affect ($F(11, 69) = 2.08$, $p < .001$; $r^2 = .25$, adjusted $r^2 = .13$). Specifically, participating in groups that displayed Working Together and Edgy Compliance dynamics predicted greater anticipated positive affect reported by students at end of the study ($\beta = .43$, $p = .004$, and $\beta = .30$, $p = .041$, respectively). Small-group dynamics did not predict anticipated negative affect.

4.1.3 Alignment between self-report and observational group data

To examine the alignment between student and researcher perceptions, we (a) created group-averaged How I Was scores for the same lessons for which we had coded data and then (b) examined the associations between self-reported group behavior and observed group behavior. Group is the unit of analysis ($N = 24$).

Group-averaged reported Enhancing behavior positively associated with Working Together and Resource Drain co-regulation dynamics ($r = .27$, $p = .010$; $r = .31$, $p = .003$, respectively). Group-averaged reported Withdrawn behavior positively associated with both Resource Drain ($r = .24$, $p = .022$) and Conflict and Control ($r = .23$, $p = .027$) dynamics. Finally, group-averaged reported Withdrawn behavior also negatively associated with the Working Together dynamic ($r = -.29$, $p = .004$).

4.2 Interpretations

Fusing the self-report and real-time observation data provided two main insights into students' experiences in small groups. First, results indicated that students were aware of themselves within their groups. The self-reported small group behavior aligned with the systematic observation data. For example, self-reported Enhancing behavior positively associated with Working Together co-regulation, while self-reported Withdrawn behavior associated with greater Conflict and Control and less Working Together co-regulation. Further underscoring the alignment between the two data sources, Resource Drain co-regulation—dynamics in which group members expressed needs (e.g., by asking for materials, attention, etc.)—associated with both self-reported Enhancing and Withdrawn behavior. In this instance, groups consisted of members who provided help (Enhancing) to those who needed and/or wanted it (Withdrawn). Thus, students as young as grade three appear to accurately self-monitor, and students as old as grade five appear willing to accurately report their small-group behavior.

Second, results indicated that students' emotional adaptation and academic readiness were important features of small-group dynamics and personal learning. For example, participation in Edgy Compliance co-regulation dynamics, unpleasant as it may have been, predicted an increase in students' posttest endorsement of the Regret and Repair emotional adaptation profile. This suggested that students were not only aware of the behavior exhibited in their small group but that they learned about themselves and others from that experience. In this instance, interpersonal dynamics informed intrapersonal endorsements that appeared to move the student away from their prior experience toward the person they wanted to become—the person who feels badly when failing to support another and works to make amends.

Students' academic readiness also provided evidence for the press between intra- and interpersonal dynamics. Groups with higher-ranked students, for example, were more likely to display Working Together and Edgy Compliance co-regulation dynamics. This suggests that students with higher math readiness have the potential to direct their resources in more productive (e.g., offering suggestions, asking questions) and less productive (e.g., bragging, refusing others' participation) ways. Yet, in spite of the different real-time

interaction patterns, the self-report data indicated that students learned from these experiences and that the small groups shaped students' posttest characteristics. As noted with Edgy Compliance co-regulation, these dynamics predicted greater endorsement of positive regulation strategies moving forward (Regret and Repair); and, for the groups already Working Together, these dynamics predicted lower anxiety at posttest.

5. Discussion

In this paper, we addressed two of the three guiding questions in this special issue: How does the use of self-report constrain the analytical choices made with self-report data, and how do the interpretations of self-report data influence interpretations of findings? Our goal was to illustrate the benefits and challenges of using self-report data to understand students' experiences in and attitudes towards small groups.

Overall, using self-report data alone provided insight into students' self-awareness and perceptions of their own behavior during small group. While useful for capturing the perceived student experience, relying purely on self-report data constrained our inquiry to individual-level analyses in ways that ignored the mutual give-and-take between the individual and their group members. Thus, the primary challenge of using self-report data to study small groups is that we fail to capture the dynamic processes that are inherent in these social learning tasks. In other words, self-report data can illustrate how personal sources of influence (e.g., math readiness, emotional adaptation) shape students' experiences and emergent identities; yet, we fail to also learn how students shape—and are shaped by—the mutual press between personal and social sources of influence in real time.

Although we emphasized the role of self-report data for understanding students' experiences in small groups, this paper also identified strengths and limitations of observation-only data. For example, our real-time data corroborated research by Ladd and colleagues (2014) in which students reported the (lack of) positive small-group behaviors displayed by their peers. The researchers noted that "substantial proportions of participants received average ratings that were so low…as to imply that they "seldom" or "never" exhibited such skills during collaborative classroom activities" (p. 169). Our data provided insight into which behaviors and skills these peers might engage in instead. Working Together is great when it happens, but the pursuit of joint activity, in which disagreements are respected, questions appreciated, peer elaborations valued, and understandings deepened, does not represent the reality of the array of small group dynamics. Instead, groups also display aggressive and regressive coping behaviors that result in more or less effective group functioning.

Yet, while observation-only data yielded a broader understanding of real-time behavior in small groups, we nevertheless failed to capture students' perceived experiences within these dynamics. Rather, by fusing self-report *and* observational data, we learned that small-group co-regulation dynamics were saturated with social and self-conscious emotions, and the uneven regulation of those emotions often did not proceed smoothly or turn out well. Overall, students differed in their typical need to cope with learning difficulty, but coping with lack of control and uncertainty are part of what it means to be in a small group for most. To be in a small group with peers—classmates who vary in their own learning and social skills (Ladd et al., 2014; Rogat, et al., 2013)—can exacerbate or attenuate that reality. Thus, like others in this special issue (e.g., Rogiers, et al, 2020; van Halem, et al., 2020), we argue that using multiple data sources can yield broader understandings of student behavior in classroom settings.

5.1 Considerations

The focus on self-report data in this paper and special issue warrants further discussion of survey data in particular. First, some critiques of self-report data question whether participants can provide accurate responses to researchers' survey items. These concerns broadly reflect literature showing that participants sometimes fail to understand their own motives (Nisbett & Wilson, 1977) or provide opinions about events that did not happen (Bishop, et al., 1980). However, the present study provides evidence that students in grades three and five can (and are willing to) provide accurate reports of their time in small groups. Students' self-reported individual group behavior aligned with researcher-observed behaviors taking place during the small-

group activities. For researchers, this suggests the utility of using self-report data in research on small-group processes, particularly when these survey measures can be corroborated with other data sources. Furthermore, in line with prior recommendations (Corno, 2011), teachers may also want to consider using brief surveys to better understand how small-group activities unfolded in their classrooms. Of course, while this strategy may help teachers to refine these activities in their classes, future research will also need to determine whether students' responses vary depending on whether students are reporting co-regulation dynamics to their teachers or to researchers.

Second, in addition to considering participants' understanding of their own attitudes, some researchers express concerns about using self-report data due to the surveys themselves. These criticisms acknowledge that there may be aspects of any given survey that can prevent participants from responding as accurately as possible (Duckworth & Yeager, 2015). In the current paper, the internal consistency reliability coefficients generally provided one source of evidence for using these survey measures in our analyses. However, it is important to acknowledge that one of our five *School Situations* factors, Minimize and Move On, fell below the recommended .70 for Cronbach's alpha (Nunnally, 1978). Thus, even though this factor—capturing student escape strategies—was identified in previous research using this same instrument (McCaslin, et al., 2016), we encourage researchers to replicate this work to determine whether the five School Situations factors emerge in their own samples, or whether some strategies do not translate across all contexts. In spite of the relatively low internal consistency for the Minimize and Move On factor, we are confident in our self-report measures, again due to the alignment between the survey and observational data in this study.

5.2 Conclusion

In sum, conceptions of small-group members in terms of their cooperative or regulatory skill set is a start that is likely to sputter without recognition of the fullness of individuals who have personal histories and concerns that make them more and less vulnerable to threat (Frijda, 2008) and making threats. This calls for expanding conceptions of small-group cooperative skills and dispositions of individuals to include, for example, considerations of power and influence among group members. Our observations of demanding behavior and provocative exchange suggest that students can consider power from a perspective of coercion and control rather than one of support and positive influence (Keltner, 2016). Students' personal concerns and heightened perceptions of threat are part of power and influence dynamics. Both are better understood within deliberate consideration of conflicts that may underlie and result from them. We do students a disservice when we fail to acknowledge the fullness of the task of working and learning with others. We also miss an opportunity to fully learn from the potential of small-group learning for students' personal growth and well-being when we fail to use multiple research methods fluidly. Thus, while self-report and observational data alone can each increase our understanding of student motivation and learning processes, pursuing both in tandem can yield richer understandings of students' classroom activity, how these experiences evolve over time, and how that matters in the dynamics of being and becoming a student.

Keypoints

- Students' self-reported behavior during small group predicted their reported end-of-study anxiety and anticipated emotion.
- Real-time audio data indicated five distinct types of co-regulation dynamics that students can engage in within small groups (e.g., communication patterns, coping, etc.).
- Students' initial group-level characteristics predicted their real-time co-regulation dynamics.
- Co-regulation dynamics during small group predicted individual students' self-reported end-of-study anxiety, anticipated emotion, and emotional adaptation.
- The real-time audio data corroborated students' self-reported behavior during small group.

References

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-opinions on public affairs. The Public Opinion Quarterly, 44(2), 198-209.

Burggraf, S. A. (1993). School situations. Unpublished manuscript. Bryn Mawr, PA: Bryn Mawr College.

Corno, L. (2011). Studying self-regulation habits. In H. D. Schunk, & B. Zimmerman (Eds.), Handbook of self-regulation of learning and performance (pp. 361-375). New York: Routledge.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for education purposes. Educational Researcher, 44(4), 237-251. https://doi.org/10.3102/0013189X15584327

Elias, M. J., & Schwab, Y. (2006). From compliance to responsibility: Social and emotional learning and classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), Handbook of classroom management: Research, practice, and contemporary issues (pp. 309-341). Mahwah, NJ: Lawrence Erlbaum Associates.

Frijda, N. H. (2008). The psychologists' point of view. In M. Lewis, J. M., Haviland-Jones, & L. F. Barrett (Eds.), Handbook of emotions, 3rd ed. (pp. 68-87). New York: Guilford Press.

Fryer, L. K., & Dinsmore, D. L. (2020). The promise and pitfalls of self-report: Development, research
design and analysis issues, and multiple methods. *Frontline Learning Research*, *8*(3), 1–9. http://doi.org/10.14786/flr.v8i3.623

Hadwin, A. F., & Järvelä, S. (2011). Introduction to a special issue on social aspects of self-regulated learning: Where social and self meet in the strategic regulation of learning. Teachers College Record, 113(2), 235-239.

Hadwin, A. F., Järvelä, S., & Miller, M. (2018). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In D. H. Schunk & J. A. Greene (Eds.), Handbook of self-regulation of learning and performance (pp. 83-06). New York, NY: Routledge.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. Educational Researcher, 38, 365-379. https://doi.org/10.3102/0013189X09339057

Keltner, D. (2016). The power of paradox: How we gain and lose influence. New York, NY: Penguin Press.

Ladd, G. W., Kochenderfer-Ladd, B., Visconti, K. J., Ettekal, I. Sechler, C. M., & Cortes, K. I. (2014). Grade-school children's social collaborative skills: Links with partner preference and achievement. American Educational Research Journal, 51(1), 152-183. https://doi.org/10.3102/0002831213507327

McCaslin, M. (2009). Co-regulation of student motivation and emergent identity. Educational Psychologist, 44(2), 137-146. https:// doi.org/10.1080/00461520902832384

McCaslin, M., & Burross, H. (2008). Student motivational dynamics. Teachers College Record, 110(11), 2319-2340.

McCaslin, M., Tuck, D., Waird, A., Brown, B., LaPage, J., & Pyle, J. (1994). Gender composition and small-group learning in fourth-grade mathematics. *Elementary School Journal, 94*, 467-482.

McCaslin, M., & Vega, R. I. (2013). Peer co-regulated learning, emotion, and coping in small-group learning. In S. Phillipson, K. Y. L. Ku, S. N. Phillipson (Eds.), *Constructing educational achievement: A sociocultural perspective* (pp. 118-135). New York, NY: Routledge.

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 140 - 163
ISSN 2295-3159

Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data

Nicolette van Halem[a], Chris van Klaveren[a], Hendrik Drachsler[bc], Marcel Schmitz[d], Ilja Cornelisz[a]

[a]Vrije Universiteit Amsterdam, the Netherlands
[b]Open Universiteit, the Netherlands
[c]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany
[d]Hogeschool Zuyd, the Netherlands

Article received 29 May 2019 / revised 26 July / accepted 17 September / available online 30 march

Abstract

For decades, self-report instruments – which rely heavily on students' perceptions and beliefs – have been the dominant way of measuring motivation and strategy use. Event-based measures based on online trace data arguably has the potential to remove analytical restrictions of self-report measures. The purpose of this study is therefore to triangulate constructs suggested in theory and measured using self-reported data with revealed online traces of learning behaviour. The results show that online trace data of learning behaviour are complementary to self-reports, as they explained a unique proportion of variance in student academic performance. The results also reveal that self-reports explain more variance in online learning behaviour of prior weeks than variance in learning behaviour in succeeding weeks. Student motivation is, however, to a lesser extent captured with online trace data, likely because of its covert nature. In that respect, it is of importance to recognize the crucial role of self-reports in capturing student learning holistically. This manuscript is 'frontline' in the sense that event-based measurement methodologies with online trace data are relatively unexplored. The comparison with self-report data made in this manuscript sheds new light on the added values of innovative and traditional methods of measuring motivation and strategy use.

Keywords: Self-Regulated Learning; Self-Report Measures; Event-Based Measures; Online Trace Data

Corresponding author: *Nicolette van Halem, van der Boechorstraat 7, 1081 BT Amsterdam, the Netherlands,* n.van.halem@vu.nl DOI 10.14786/flr.v8i3.497

1. Introduction

Motivation and strategy use are core concepts in the literature on learning and instruction. Widely known and prominent in contemporary educational psychology is the theory of Self-Regulated Learning (SRL), which integrates these constructs in explaining student success. SRL can be defined as "an active, constructive process of goal setting and attempting to monitor, regulate, and control cognition, motivation, and behaviour, guided and constrained by goals and the contextual features in the environment" (Dinsmore, et al., 2008; Jupp, 2006; Panadero, et al., 2016; Panadero, 2017; Pintrich, 2000, p. 453). SRL is an internal process that we cannot directly access, such that proxies are necessary to assess this SRL process (Boekaerts & Corno, 2005).

For decades, it has been argued that aptitude-based self-report instruments – which rely heavily on students' perceptions and beliefs – do not fully capture SRL. Theories of SRL emphasize that each instance of self-regulation is a function of the individual's dynamic interaction with the learning environment, but few instruments satisfactorily capture such data (Boekaerts, et al., 2000; Efklides, 2011; Veenman, 2011; Winne & Perry 2000). Yet, self-reports have remained the dominant way of measuring SRL (Boekaerts & Corno 2005; Winne and Perry 2000), as the implementation of more time-intensive data collection methods, such as thinking-aloud protocols, event-based self-reports, or observations, are often times not feasible in educational settings. The recent introduction of tracing methods in online learning environments mainly through learning analytics (Greller & Drachsler, 2012) sparked the development of an alternative event-based measurement method of SRL, enabling a form of online observation methods, while influencing the learning process as little as possible (Panadero et al., 2016). These measurement methods are spurred by active efforts in the learning analytics community to bridge the gap between learning sciences and data analytics. However, so far learning theories such as SRL are seldom used as theoretical basis for the design and evaluation of tracing methods (Jivet, et al., 2017; Jivet, et al., 2018). Furthermore, there is a dearth of empirical work into the potential of online observation methods to complement self-reports on SRL. The purpose of this study is therefore to triangulate constructs suggested in SRL theory and measured using aptitude-based self-reported data with revealed online traces of learning behaviour.

This study takes place in the context of the implementation of a Learning Analytics application, called 'the Learning Analytics Experiment' in Dutch higher education. The main aim of the project was to create an opportunity for institutions, teachers, and students to gain experience with different facets of learning analytics (e.g. privacy, feedback provision, insight in the use of learning materials, etc.). The online traces of students' learning behaviour were recorded using xAPI (or Tin Can) trackers, which have the potential to track learning experiences and store records of learners' (e.g. '*access video'* or '*receive grade*'). The trackers used in this study captured information specifically on the use of learning materials, such that the use of the online learning environment can be compared between students over time in light of different components of SRL. This study describes an implementation of the xAPI trackers in a mandatory first-year statistics course at a Dutch university, during two consecutive academic years. During the implementation, self-reports on motivation and strategy use were collected to triangulate the trace data. This offered a rich case to unpack the aggregated data collected with self-reports and to, vice versa, colour the trace data with self-reports on motivation and strategy use. Accordingly, this study aims to shed light on two of the central questions addressed in this special issue: *In what ways do self-report instruments reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use?* And: *How does the use of self-report constrain the analytical choices made with that self-report data?*

2. Theoretical framework

2.1 Complementarities of self-reported SRL measures and online trace data

Online learning environments are central in today's higher education, as they not only form a learning portal for a variety of purposefully selected learning resources, but also help to navigate through the course,

enable students to be in contact with the instructor and peers, and to engage in various learning activities in a student-led fashion (Lust et al., 2012; Molenda, 2008). Students leave traces when interacting with these online learning environments. Trace data can be defined as "observable evidence of particular cognitions that are obtained at points where a cognitive process is applied while completing a task" (Howard-Rose & Winne, 1993, p. 594). A growing body of literature confirms the importance for online learning environments for learning outcomes. Gašević, et al., (2014), for example, showed that the number of logins, number of operations performed on discussion forums and resources accounted for approximately 21% of the variance in academic performance. This finding is in line with earlier research on the relation between the frequency of visits to an online learning environments and students' academic performance (Coogan et al., 2005; DeNeui & Dodge, 2006; Wang & Newlin, 2000). Models on SRL provide a holistic theoretical foundation for the relation between observed student behaviour in online learning environments and academic performance, based on the cognitive, metacognitive, behavioural, motivational, and affective aspects of learning (Panadero, 2017). One of the latest meta-analyses on the effect of SRL (Sitzmann and Ely, 2011) shows that the four biggest predictors of learning gains — goal level, persistence, effort, and self-efficacy — have a significant motivational value. Longitudinal investigation of the interaction between motivation, strategy use, and the learning environment is, however, scarce (Panadero, 2017).

There are several reasons to believe that online trace data hold potential for tapping into the process of SRL. Students (especially in higher education) are the agents in online environment usage: they determine which resources are used and how these resources are used. The effects of the instructional design of a learner's environment are therefore never deterministic, since the use of the environment depends on the personal goals, motivation, and volition of the student (Winne & Baker, 2013). Usage of online environment can be considered a skill in itself, as it requires a repertoire of learning strategies, confidence, and competences. As Lust and colleagues (2012) put it, online environments are only beneficial when learners recognize the learning resources as a learning opportunity for which they are motivated to spent effort and time on. In other words, effective use of online learning environments can be conceptualized as a manifestation of SRL. The extent to which an individual student is motivated and able to self-regulate their learning process is thus a prerequisite for effective tool-use (Winne & Baker, 2013; Lust, et al., 2012). In addition, and elaborately described by Fryer (2017), the relation between motivation, strategy use, and the use of learning environments can be conceptualized as reciprocal. Namely, learning activities undertaken in the process phase of learning, as well as the resulting product of learning, feed back to students' beliefs, attitudes, and ideas around motivation, strategy use, and self-regulation that play a role in the presage stage of learning. Online trace data, thus, reflects the dynamic relation between students and their learning experiences over time.

2.2 Removing analytical restrictions of inventories on motivation and strategy use with online trace data

An event-based measure of SRL with online trace data arguably has the potential to remove analytical restrictions of aptitude measures. Firstly, and according to the overconfidence effect, unrealistically favourable attitudes that people have towards themselves (Taylor & Brown, 1988) can impose an upward bias in self-reports of motivation and strategy use. In other words, students might apply less and less effective study strategies than self-reported. Zhou and Winne (2012) confirm this theory empirically by showing that "trace data"-based measures of student achievement goal orientation were much stronger associated with learning outcomes than with self-reported ones. This is particularly pressing since existing research suggests that learners tend to use ineffective learning strategies (Jamieson-Noel & Winne, 2003), and do not make effective use of available resources to optimize their learning, even in those environments that build on effective learning designs (Ellis, et al., 2005; Lust, et al., 2013). As discussed in this special issue, potential reasons are students' cognitive processing capacity (Chauliac, et al., 2020), exerted effort (Iaconelli & Wolters, 2020), and student characteristics (Vriesema & McCaslin, 2020). Comparing self-reported data with online traces of learning behaviour taps right into a potential upward bias in self-reports on motivation and strategy use. Since online traces of learning behaviour do not suffer from socially desirable responding bias, it is possible that they provide a better approximation of motivation and strategy use. Secondly, inventories are restricted as they often predetermine the level of aggregation in the analysis of data on motivation and strategy use (e.g. fixed at student-level), which, as a result, generates student-focussed or aptitude-based measures. This jeopardizes the

potential of adapting effectively to individual needs and preferences during a study episode or educational program. Alongside self-report instruments that operationalized motivation and self-regulation as event-bound (Winne, 2010), event-based trace data of learning behaviour can provide a dynamic insight in how motivation and strategy use does not only vary between students, but also within students over time. This is particularly relevant since previous research show that self-motivational beliefs and strategy use can vary considerably over the course of a study episode, or throughout the educational program (Boekaerts et al., 2000; Efklides, 2011; Veenman, 2011; Winne & Perry 2000). One example is presented in this special issue by Moeller, et al., (2020) in their study on this intra-individual variation in self-motivational beliefs.

2.3 Status of research on SRL combining self-reports and online trace data

Up until now, empirical studies on SRL that use online trace data often aim at identifying different patterns of usage behaviour. This has led to a wide variety of student typologies aiming at a better understanding of the type of learning strategies used by students, looking at the use of information-tools (such as forums, instruction video's, and interactive mind maps) over the duration of a course (Heffner & Cohen, 2005; Hoskins & van Hooff, 2005; Huon, et al., 2007). Examples of typologies are the active and passive users of forums (Hoskins & van Hooff, 2005); the early-, constant-, and late users (Knight, 2010; Lust, et al., 2012); the low-, average-, and high frequency users (Bera & Liu, 2006; Jiang, et al., 2009). Research methods in identifying learning strategies inferred from online trace data evolve rapidly. A recent strand of research on trace data adopts data mining techniques in detecting striking, previously unknown, patterns in study tactics and strategy use (Han, et al., 2011). In general, it remains a challenge to qualify these patterns, typologies, and clusters of study tactics, and to make these insights actionable for instructional design accordingly. So far, studies repeatedly find that usage patterns do explain variance in academic performance (e.g. Cho & Yoo, 2017; Cornelisz & van Klaveren, 2018; Gašević et al., 2014; Han et al., 2011; Schmitz et al., 2018), but what type of self-regulation it represents and the quality thereof remains a black box. Only a handful of studies focussed on triangulation of self-reports specifically on SRL using online trace data of student learning. The few studies that describe a quantitative comparison of self-reports on SRL and online trace data show that the relation between these two sources of data is not straightforward. Hadwin, et al. (2007) used ten relevant items from the motivated Strategies for Learning Questionnaire (MSLQ) on strategy use to explain students' learning in gStudy, a web-based learning environment in which students read texts, summarize, and use concept maps in an introductory educational psychology course (N = 188). They clustered students into four groups based on self-reported SRL data and found their actual learning patterns in gStudy differed substantially, even among students from the same cluster. Kim, et al. (2018) made a similar comparison among undergraduates in Korea. They analysed online trace data from 284 undergraduate students enrolled in an asynchronous online statistics course. Based on self-reports collected with the MSLQ, students were classified as fully, partially, or not self-regulating. Surprisingly, this distinction did not reveal different patterns in online traces of learning. The main difference between the groups was timing in study behaviour, students classified as not self-regulating studied mainly shortly before the examination, which was negatively related to academic performance. A study of Guerra, et al. (2016) used the achievement-goal questionnaire and online trace data (N = 89). Their study shows that students who report a high mastery-approach show a higher level of activity in the online learning environment. There results also suggest that highly motivated students are more sequential in their patterns of navigation, which means students are less likely to follow the suggested order of topics to study. Cho & Yoo (2017) adopted a different approach and compared precision in prediction of students' academic performance based on patterns in self-reports based on MSLQ and patterns in online trace data established with data mining techniques (N = 60). The model based on online trace data provided a more precise prediction. The authors note that this might be partially due to the fact that the trace data provided more variables and was based on a bigger data set. Also, the study did not address whether or not it is likely that the self-reports and online trace data actually measured the same constructs. Like in the studies described earlier, the inventory is aptitude-based (Muis, et al. 2007; Zimmerman, 2008), whereas the online trace data is event-based. In this special issue, Rogiers et al. (2020) present one of the first studies in this space that compares event-based self-report data with trace data. Overall, the interpretation of online trace data in light of self-reports is not clear-cut, there are

many questions left unanswered about the relationships between the constructs measured with an instrument such as the MSLQ and online trace data collected in different education settings.

3. Research questions and hypotheses

The aim of this study is to further explore the relationship between measures of SRL through self-reports and online trace data, by scrutinizing variance in self-reports and online trace data between and within students. Because the online traces of learning behaviour are event-based, our study provides a dynamic insight in how motivation and strategy use vary between *and within* students over time. The findings are instrumental in guiding innovations in education towards effective personalized learning. Accordingly, the following research questions are formulated for this study:

1. To what extent do self-reports explain variance in online trace data of learning behaviour?
2. How stable is the relation between self-reports and online trace data throughout the various weeks of the course?
3. How well do self-reports explain student performance in comparison to online trace data?

Following Fryer (2017), we expect that self-reports on motivation, strategy use, and self-regulation gauge between-student differences in the presage stage and that online trace data gauges the process phase of learning. With respect to the first research question, it can, therefore, be expected that self-reports explain substantial variance in learning behaviour observed through online trace data. Furthermore, student differences in the presage stage will likely be affected over time by the feedback loop between the presage, process, and product phase. Therefore, it can be expected that the self-reports predict online learning behaviour most precisely at the time the self-reports are administered. With respect to research question two, thus, we expect that the relationship between self-reports and online trace data varies over time. Finally, in light of the third research question, we expect online trace data to explain an equal amount or more variation in student performance than self-reports, in line with the findings of Cho and Yoo (2017).

This study adds to the existing literature as follows. Firstly, there are only a few studies so far that tapped into the relationship between self-reports and online trace data. Given the sensitivity of the usage patterns to the instructional design of an online learning environment (Gašević, et al., 2016), it is of great importance that a broad range of educational settings are explored with online trace data. The course, instructional design, and educational setting investigated in this study is considered particularly relevant because it is highly representative for other courses in mainstream Dutch higher education and since this particular course plays a crucial role in all social science programs in the Netherlands. In addition, this study compares multiple cohorts, which yields insight in continuity and change of strategy use with the constant evolvement of course design. Secondly, the few studies that have been comparing self-reports and online trace data so far dealt with relatively small sample sizes (Cho & Yoo, 2017; Guerra et al., 2016; Hadwin et al., 2007; Kim et al., 2018) or did not measure the full range of self-motivational beliefs and strategies with self-reports (Hadwin et al., 2007; Cho & Yoo, 2017). As a result, there is no clarity, yet, on the relevance and actionability of online trace data in comparison to inventories on motivation, strategy use, and self-regulation.

4. Methodology

4.1 Participants

The data used consist of self-reports and detailed log-data on SRL, collected among two cohorts of first year students at the Faculty of Behavioural and Movement Sciences during a mandatory statistics course that took place between October and December 2016 ($N = 435$; 94.44% female, $M_{age} = 20.60$ years, $SD = 5.18$) and 2017 ($N = 489$; 78.90% female, $M_{age} = 20.88$ years, $SD = 3.48$). The 2017 cohort included international

students, as this was the first year that this course was also taught in English. The Faculty of Behavioural and Movement Sciences offers the following educational programs: movement sciences, education sciences and behavioural, developmental and clinical psychology.

4.2 Course design

4.2.1 Online learning environment

The online learning environment was available to all students throughout the course and its usage was not mandatory in any sense. In 2016, the online learning environment consisted of a learning management system (Blackboard) and a separate online learning tool, called I Hate Statistics[1]. Blackboard is one of the leading commercial LMS software packages used by North American and European universities (Itmazi & Megias, 2005; Munoz & Duzer, 2005). From the start of the academic year 2017-2018, the institution switched to a new learning management system, called Canvas. Together with the standard features of LMSs, Canvas provides advanced options like learning outcomes, peer review, migration tools, e-portfolios, screen sharing and video chat etc. Canvas is gaining popularity, hundreds of colleges, universities, and school districts currently use this package (www.instructure.com).

Across cohorts, the learning management system was structured in a similar fashion. Both Blackboard and Canvas provided students with three types of tools: 1) an information tool (lecture slides, instruction video's, and general course information), which provided the course content in a structured way; 2) a cognitive tool for self-assessments that enabled students to interact with the subject matter, to assess and to reflect on the learning content; and 3) a communication tool (forums) that enabled students to communicate with peers and instructors (Lust, et al., 2012). The tools were structured based on the week-topics of the course and were available at all times throughout the course.

The learning management system referred students to 'I Hate Statistics', which provided students with an online environment for practicing and studying, where students could engage in lessons, challenges, and self-assessments that were related to the week topics, or to other topics that were available in this environment. Each challenge consisted of on average maximal ten questions; yet, the challenge was automatically finished when students correctly answered five questions within the challenge. A unique feature of the environment was that it is built around the statistics course and offered content in a particular week that was similar to the content offered in the lectures and seminars.

The xAPI tracking method is applicable for all online learning environments (see for example https://xapi.com/). In the context of this study, it was used to gain insight into the use of learning materials in the learning management systems. The teacher created so-called 'recipes' and placed them in Blackboard, where the use of information resources and participation in online activities were tracked. These recipes ensured that the desired statistics were generated, as well as information on the type of tool that was used (e.g. slides, challenges, lessons) an action verb (e.g. accessed, received), and a label (e.g. lessons on the chi-square test, lecture slides of week 1). A comparable tracking method was used in I Hate Statistics, which generated similar data about the use of the learning materials.

4.2.2 Instructional design

During the eight-week introductory statistics course, students had to attend two lectures each week in which theoretical concepts were addressed. Additionally, they had to attend one seminar each week with mandatory attendance in which the assignments and the subject matter were discussed and opportunities for peer- and teacher feedback were organized. Offline and in the course manual, students were referred to the textbook that the teacher selected as a starting point for the course. Use of this textbook is not traced within the online learning environment. The online learning environment provided opportunities for self-assessments. The self-assessments in the learning management systems and *I Hate Statistics* were similar in nature and contained

multiple-choice questions in which knowledge and comprehension were assessed. The learning management system contained four self-assessments; *I Hate Statistics* contained eight self-assessments on the course topics. In the second year of this study, it also gave access to a practice exam, with questions that were representative for the final exam. At the end of the course (week 8) each student was graded based on a final multiple-choice exam and on a research report.

4.3 Instruments

4.3.1 Self-reports

The motivated Strategies for Learning Questionnaire (MSLQ), a measure developed by Pintrich and colleagues (McKeachie, et al., 1985; Pintrich, 1991; Pintrich, et al., 1987; see also Duncan and McKeachie, 2005, for a more in-depth discussion), was used to assess self-reports on SRL. The MSLQ was derived from an extensive body of literature and was one of the first inventory on the quality of student learning that not only included attitudes, motivation, and strategy use, but also self-regulatory strategies (Entwistle & McCune, 2004). The MSLQ was progressive at the time by including the dimension of students' consciousness about the teaching-learning environment, leading to adaptation of ways to tackle academic work (Entwistle & McCune, 2004). Several studies argue that there is piecemeal evidence for the scale structure of the MSLQ (Hilpert, et al., 2013; Tock & Moxley, 2017), yet, to this date, this inventory for students in higher education is still considered relevant in light of the wide range of motivation, affect, strategy use, and self-regulation it covers. Appendix A provides a description of the scales, along with a couple of sample items per scale.

The MSLQ was administered once in the seminar of week four, along with several general questions about background variables, such as age and gender. The MSLQ contained 81 questions of which 31 items determined students' motivational orientation towards the course and 50 items assessed metacognition. The questions assess the propensity of students to engage in self-regulated learning within the specific context of this course, but, overall, the MSLQ has been classified as an aptitude measure of self-regulated learning (Muis et al., 2007; Zimmerman, 2008). Students answered with a 7-points Likert scale, ranging from 'not at all true for me' to 'very true for me'. The motivational orientation is divided into six subscales: Intrinsic goal orientation, extrinsic goal orientation, task value, self-efficacy, control beliefs and test anxiety. Metacognition was scored on nine subscales: rehearsal, elaboration, organization, critical thinking, metacognitive self-regulation, time and study environment, effort regulation, peer learning and help seeking. A definition per subscale is provided in Appendix A. For a complete description of the MSLQ and each of its subscales we refer to the manual of the MSQL (Pintrich, 1991).

Reliability coefficients were determined with the Cronbach's Alpha (see Appendix A). It is important to note that the reliability of the goal orientation sub-scales is poor. Scrutinizing the data per item did not point out particular weak items that could be removed from the scale to improve reliability.

4.3.2 Online trace data

Online trace data was obtained as part of the project 'the Learning Analytics Experiment'. The teacher of the course was actively involved in defining what type of data was collected. The teacher was facilitated to track any learning activity in the learning management system with trackers designed by SURFnet and the Amsterdam Center for Learning Analytics (ACLA). The trackers were based on xAPI recipes (a set of rules). The teacher defined the set of rules, based on activities, verbs, and labels. For example, 'formative exam' (activity) - 'accessed' (verb) - 'week 3' (label). After defining a recipe, a HTML-code was provided that accordingly was embedded in the online learning environment, often in the form of an empty object in the environment.

In addition, the designers of the application I Hate Statistics provided access to the data they collected on each learning activity a student engaged in, as well as the timestamp, the length of the learning activity, the

number of questions that a student answered within a lesson or a challenge, and the success rate within a lesson or challenge was logged for each student.

After processing the data, variables were selected to be included in the current study, following the work of Guerra and colleagues (2016), Kim and colleagues (2018), and Theobold and colleagues (2018), These variables are described in Table 1. The count- and time- based variables were aggregated to the week level based on the week-based structure of the course; Each week was structured around a particular topic. The variables Sequential Navigation and Distributed Learning provide a metric (respectively a ratio and a count measure) on student level, in order to deduce an overall measure of students' patterns in use of the online learning environment throughout the eight weeks of the course.

Table 1

Interpretation of Online Trace Data in Week- and Student Level Variables

Variable	**Operationalization**
Total Time Investment in Practicing	Week level count of minutes of practicing time in I Hate Statistics. I Hate Statistics provided the only online opportunities to practice the subject matter, time spent practicing was captured by the application. The xAPI trackers on the other resources available online did not capture the time students spend interacting with it, which means this is excluded from this variable.
Total Participation in Learning Activities	Week-level count of the number lessons, challenges, videos, lecture slides, course information, and blog posts students accessed during the eight-week course. Self-tests and forum visits are excluded from this variable.
Seeking and Providing Help	Week-level count of the number of forums threads accessed by the students during the eight-week course.
Self-Assessment	Week-level count of self-tests students took. In the first four weeks of the course, two self-tests were available to the students. In the sixth week, a practice exam was available to the students online. Self- tests contained several multiple-choice questions or open questions that covered different topics of the lecture. At the end of each self-test, students received feedback and an explanation if an answer was wrong. Each student had permission to run every self-test as many times as desired.
Sequential Navigation	Week-level count of within-topic or next-topic accessed resources. Every act of accessing a learning resources (as defined under total participation in learning activities) was classified into four different groups: (1) *within-topic*, which indicates accessing a learning activity related to the week topic; (2) *next-topic*, which indicates accessing a learning activity related to next week's topic (according to the sequence of topics in the course), (3) *jump- forward*, which indicates accessing a learning activity related to a topic of two or more weeks away from the current topic, and (4) *jump- backward*, which indicates accessing a learning activity related to a previous week topic. A *within-topic* or *next-topic* move represents a sequential navigation, while a *jump-forward* or *jump-backward* move represents a non-sequential navigation pattern. We then computed the ratio of sequential navigation versus non-sequential accessing acts per student per week up until week 6. In each of the first 6 weeks a new topic was introduced, week 7 covered all topics and week 8 was dedicated to the final exam.
Distributed Learning	Student level count of the number of weeks in which each student had accessed learning resources in the online learning environments irrespective of the actual amount of time students spent online. Higher values on this variable suggest a more distributed, continual engagement with the course content. Values on this variable could possibly range from zero to eight, as the course duration was eight weeks.

4.3.3 Academic performance

Students' academic performance is measured with the summative course evaluation. At the end of the course students' final grade was determined based on the results of a multiple-choice exam and the grading of the research report. The exam included 30 four-answer choice questions; ten questions targeted knowledge, ten targeted insight, and the other ten targeted calculations. Appendix B provides sample questions of each category. The exam was designed by the course coordinator. In general, all course coordinators are equipped with a training that covers basic knowledge and skills on constructing multiple choice tests. The course coordinator had, furthermore, access to a data base with high-quality questions used in previous years, which – albeit slightly modified – could be used to put together the exam. The exam underwent peer review and psychometric tests, the latter was used in the grading process. Final grades were scored on a scale from 1 to 10 with 10 being the highest and with 5.5 as passing threshold.

4.4 Procedure

Students were provided with an informed consent form for both the self-report data and the online trace data. In the 2016 cohort, informed consent for the online trace data was obtained at the beginning of the course, whereas informed consent for the self-report data was obtained during the student survey in week 4 of the course. In the 2017 cohort, students were provided with one informed consent form including permission to use both sources of data at the beginning of the course.

4.5 Analysis strategy

Data preparation procedures were applied prior to the analysis. Homogeneity of the data across cohorts was established for motivation and strategy use reported by students, for the variables based on the online trace data, and for academic performance, using Levene's test for homogeneity of variance and independent-sample t-tests. Multiple imputation methods were applied to deal with missing data on items of the MSLQ scales in the following analyses, using the Markov chain Monte Carlo method with a number of 10 iterations, with IBM SPSS Statistics 25.

To answer the first research question, the extent to which student self-reports tap motivation and strategy use as reflected in the online trace data online learning behaviour was examined. To that end, explained variance in online trace data of learning behaviour by student self-reports was investigated with ordinary least square regression analyses. Total participation in learning activities was used as dependent variable, because it identified between-student and between-week variance and comprised student engagement in both the learning management system and the online practicing environment. After looking at the data on the course level, models were tested on a week level, in order to gauge variability in the relation between online trace data and self-reported data based on the MSLQ and answer the second research question.

To answer the third research question, the relation between self-reports and online trace data was gauged with a third proxy of student motivation and strategy use: student academic performance based on students' grade on the multiple-choice exam at the end of the course. The proportion explained variance in academic performance by online trace data and self-reports was identified with ordinary least square analyses, separately and combined. In all analyses, students' gender and age were included as control variables.

5. Results

5.1. Descriptive statistics

In total, 605 students gave active consent to include their data collected during the course for these students could be used in the analyses. The results of the homogeneity tests and descriptive statistics are shown in Table 2. Both the self-reports and the online trace data differ significantly between the cohorts, likely related to the introduction of the practice exam for the 2017 cohort and the demographical shift based on the inclusion of international students in 2017. For example, the level of extrinsic motivation is significantly higher in the second cohort, as well as the average number of self-assessments per student per week. Even though students in the 2017 cohort engaged significantly more in self-assessments, their participation in online learning activities was on average significantly lower and they engaged in online learning activities in six out of eight weeks versus seven in 2016.

Table 2

Tests of Homogeneity

	Cohort 2016		**Cohort 2017**		**Levene's test for equality of variances**	**T-test for equality of means**
	N	***M* (*SD*)**	***N***	***M* (*SD*)**		
Age	121	20.57 (5.18)	393	20.84 (3.43)	*	
Grade Multiple Choice Exam	116	6.80 (1.86)	429	6.57 (2.31)	***	
Grade Research Report	111	6.48 (1.58)	337	6.81 (1.67)		
Intrinsic Goal Orientation	120	4.89 (0.87)	396	4.74 (0.87)		
Extrinsic Goal Orientation	120	4.09 (1.32)	396	4.49 (1.10)	**	**
Task Value	120	4.81 (0.90)	396	4.95 (0.85)		
Self-Efficacy	120	4.54 (0.91)	396	4.28 (1.05)		*
Test Anxiety	120	4.17 (1.20)	396	4.24 (1.31)		
Control of Learning Beliefs	120	5.47 (0.87)	396	5.43 (0.92)		
Metacognitive Self-Regulation	120	4.16 (0.70)	395	4.37 (0.76)		**
Rehearsal	119	4.51 (1.03)	395	4.28 (1.19)		*
Elaboration	119	4.83 (0.82)	324	5.05 (0.85)		*
Organization	119	4.50 (0.98)	395	4.64 (1.11)	*	
Critical Thinking	119	3.20 (0.87)	395	3.72 (1.14)	***	***
Time and Study Environment	119	4.58 (1.12)	395	4.76 (0.96)	*	
Effort Regulation	119	4.87 (1.14)	394	4.58 (1.24)		*
Peer Learning	119	3.60 (1.16)	395	3.70 (1.27)		
Help Seeking	119	4.52 (1.12)	395	4.06 (1.23)		**
Total Time Investment (min)	120	36.81 (76.94)	429	29.20 (98.04)		
Total Participation	121	20.51 (7.18)	479	12.48 (7.02)		***
Seeking and providing help	121	0.92 (3.00)	479	1.12 (2.26)		
Self-Assessment	121	1.25 (0.71)	479	8.18 (10.54)	***	***
Sequential Navigation	121	0.35 (0.15)	470	0.34 (0.20)	*	
Distributed Learning	121	7.19 (1.00)	479	6.20 (1.94)	***	***
						Chi-square test, χ
Gender (1 = female)	121	0.87 (0.34)	395	0.81 (0.39)		7.64, $p = .17$

* α =.05; ** α =.01; *** α =.001

5.3 The relationship between self-reports and online traces of motivation and strategy use over time

The variables obtained through self-reports were regressed on online trace-data, both on student- and week-level, using standardized values, revealing how stable relationship between self-reports and online trace data is over time. Student controls (age, gender, and cohort) were included in each of the models. The results are presented in Table 3. A comparison of the adjusted R^2 of the baseline model and the model of the sum of participation over the whole course reveals that students' self-reports on motivation and strategy use explain about 9% of variance in the resources accessed in the online learning environment.

The adjusted R^2 of the models across weeks indicate that students' self-reports seem to explain more variance in participation in the weeks prior to the collection of the self-reports (week 4) than participation in the weeks afterwards. The self-reports on time and study environment management explain a substantial proportion in variance in participation up and including week 4, but this changes in the weeks after the self-reports were collected. Self-reports on rehearsal and elaboration strategies emerge as significant predictors of participation in week 3-6 and week 8.

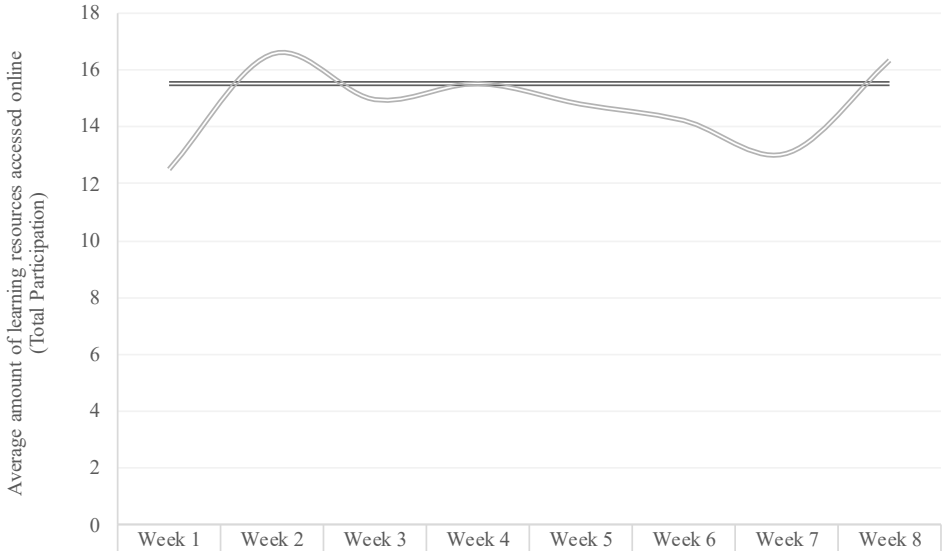
Figure 1 shows the extent to which predictions of participation based on the online trace data of week 4 and the self-reports are representative for participation in other weeks. Actual participation levels reveal that student participation fluctuates over time, with on average a peak in week 2 and week 8.

Table 3

Self-Reports Regressed on Online Trace Data

Dependent variable: Total Participation	Baseline	Sum	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**
Model 1: Self-Reports	N = 604	N = 604	N = 457	N = 520	N = 542	N = 531	N = 505	N = 460	N = 421	N = 445
Intercept	-.47 (.09)***	-.39 (.10) ***	-.47 (.10)***	-.42 (.11)***	-.22 (.09)*	-.13 (.09)	-.26 (.12)*	-.14 (.12)	-.07 (.12)	-.07 (.20)
Student controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cohort control	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intrinsic Goal Orientation	n/a	-.10 (.05)*	-.06 (.05)	-.06 (.05)	-.04 (.05)	.02 (.05)	-.04 (.06)	-.01 (.06)	.02 (.06)	-.10 (.11)
Extrinsic Goal Orientation	n/a	-.01 (.04	.03 (.04)	-.05 (.05)	.04 (.04)	-.01 (.04)	-.06 (.05)	-.01 (.05)	-.02 (.05)	.07 (.09)
Task Value	n/a	.03 (.05)	.02 (.05)	.02 (.05)	.04 (.04)	-.04 (.05)	-.04 (.06)	-.07 (.06)	.00 (.06)	.05 (.11)
Self-Efficacy	n/a	.03 (.05)	.09 (.05)	.09 (.06)	.02 (.05)	-.04 (.05)	.12 (.06)	-.10 (.07)	-.08 (.06)	.06 (.11)
Test Anxiety	n/a	-.01 (.04)	.02 (.04)	.04 (.05)	.04 (.04)	.00 (.04)	.08 (.05)	-.01 (.05)	-.02 (.05)	.04 (.09)
Control of Learning Beliefs	n/a	.06 (.04)	.04 (.04)	-.01 (.05)	.01 (.04)	.04 (.04)	.05 (.05)	.08 (.05)	.06 (.05)	-.03 (.09)
Metacognitive Self-Regulation	n/a	-.01 (.05)	-.01 (.05)	.01 (.06)	-.07 (.05)	-.06 (.05)	-.06 (.06)	.08 (.07)	.00 (.06)	.16 (.11)
Rehearsal	n/a	-.05 (.04)	-.07 (.05)	-.07 (.05)	-.02 (.04)	-.11 (.04)**	.00 (.05)	-.11 (.05)*	.04 (.06)	.14 (.09)
Elaboration	n/a	-.12 (.05)*	-.02 (.05)	-.08 (.05)	-.10 (.04)*	.06 (.05)	.05 (.06)	.10 (.06)	-.08 (.07)	-.31 (.11)**
Organization	n/a	-.02 (.05)	-.05 (.05)	.03 (.05)	.00 (.04)	.03 (.04)	.04 (.06)	-.15 (.06)	.01 (.06)	-.02 (.10)
Critical Thinking	n/a	-.02 (.04)	-.06 (.04)	-.03 (.05)	.00 (.04)	-.01 (.04)	-.03 (.05)	-.04 (.06)	-.01 (.05)	.11 (.10)
Time and Study Environment	n/a	.30 (.05)***	.31 (.05)***	.23 (.06)***	.19 (.05)***	.24 (.05)***	.11 (.06)*	.11 (.06)	.08 (.06)	-.02 (.11)
Effort Regulation	n/a	.08 (.05)	.02 (.05)	.09 (.06)	.09 (.05)	.02 (.05)	.03 (.06)	.03 (.06)	-.08 (.06)	-.13 (.12)
Peer Learning	n/a	.05 (.04)	-.04 (.04)	-.06 (.05)	-.02 (.04)	-.01 (.04)	.04 (.05)	.01 (.05)	.04 (.05)	-.01 (.09)
Help Seeking	n/a	.02 (.04)*	.08 (.04)	.05 (.04)	.03 (.04)	-.01 (.04)	.01 (.05)	.00 (.05)	-.02 (.05)	-.04 (.09)
Adjusted R^2	.19	.28	.28	.32	.21	.09	.03	.03	-.02	.04

* α =.05; ** α =.01; *** α =.001



	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
Prediction based on data collected in week 4	15.53	15.53	15.53	15.53	15.53	15.53	15.53	15.53
Actual	12.52	16.57	14.97	15.53	14.81	14.24	13.08	16.35

Figure 1. Predicted participation based on the regression of students' self-reports on total participation in week 4 and the actual participation across weeks

5.3 Explained variance in academic performance by self-reports and online traces of motivation and strategy use

The variables based on self-reports and online trace data were regressed on academic performance; using standardized values and in a stepwise fashion, see Table 4. Student controls (age, gender, and cohort) were included in each of the models. The differences in explained variance between Model 1 and Model 2 need to be interpreted with caution, as there are substantially fewer variables included in Model 2. Based on the explained variance of Model 3, it can be concluded that there is a substantial unique contribution of the self-report and online trace data-based variables in predicting academic performance. In the final model, there are three scales of the MSLQ that explain a substantial proportion in academic performance; self-efficacy, elaboration strategies, and effort regulation. Four variables based on the online trace data explain substantial variation; total time invested, total participation, self-assessment and distributed learning. The total time invested in the online learning environment is negative related to academic performance, even though the total amount of resources accessed by the student (i.e. total participation) is positively related to academic performance.

Table 4

Self-Reports and Online Trace Data Regressed on Academic Performance

Dependent variable: Grade on the multiple-choice exam	**Model 0: Baseline**	**Model 1: Self-reports**	**Model 2: Online trace data**	**Model 3: Combined**
	N	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**	***B* (*SE*)**
Intercept	604	.22 (.10)*	-.28 (.10)**	-.12 (.10)	-.21 (10)*
Student controls		✓	✓	✓	✓
Cohort control		✓	✓	✓	✓
Self-reports					
Intrinsic Goal Orientation		n/a	-.12 (.05)*	n/a	-.08 (.05)
Extrinsic Goal Orientation		n/a	-.07 (.04)	n/a	-.06 (.04)
Task Value		n/a	.01 (.05)	n/a	-.00 (.05)
Self-Efficacy		n/a	.19 (.06)***	n/a	.17 (.05)**
Test Anxiety		n/a	-.09 (.05)*	n/a	-.08 (.04)
Control of Learning Beliefs		n/a	.10 (.05)*	n/a	.08 (.04)
Metacognitive Self-Regulation		n/a	-.04 (.05)	n/a	-.02 (.05)
Rehearsal		n/a	-.07 (.05)	n/a	-.06 (.04)
Elaboration		n/a	.06 (.06)	n/a	.10 (.05)*
Organization		n/a	.07 (.05)	n/a	.06 (.05)
Critical Thinking		n/a	-.02 (.05)	n/a	-.03 (.04)
Time and Study Environment		n/a	.10 (.06)	n/a	.03 (.05
Effort Regulation		n/a	.20 (.06)***	n/a	.16 (.05)**
Peer Learning		n/a	-.00 (.05)	n/a	-.04 (.04)
Help Seeking		n/a	.02 (.04)	n/a	.03 (.04)
Online trace data					
Total Time Investment		n/a	n/a	-.12 (.04)**	-.14 (.04)***
Total Participation		n/a	n/a	.16 (.05)**	.12 (.05)*
Seeking and providing help		n/a	n/a	.05 (.04)	.02 (.04)
Self-Assessment		n/a	n/a	.24 (.05)***	.17 (.05)***
Sequential Navigation		n/a	n/a	.08 (.04)*	.01 (.04)
Distributed Learning		n/a	n/a	.08 (.05)	.10 (.48)*
Adjusted R^2		.02	.18	.18	.27

* α =.05; ** α =.01; *** α =.001

6. Conclusion and discussion

Starting from the central questions of this special issue, this study aimed to provide insight in the ways in which self-reports reflect the conceptualizations of the constructs suggested in theory related to motivation or strategy use. To that end, this study looked into the relationship between measures of SRL through aptitude-based self-reports and event-based online trace data. Using event-based measurement methods of SRL based on online trace data complementary to self-report instruments can be a first step in capturing self-regulation as a function of the individual's dynamic interaction with the learning environment (Boekaerts et al., 2000; Efklides, 2011; Veenman, 2011; Winne & Perry 2000). Capturing the individual's dynamic interaction with

the learning environment can be instrumental in guiding innovations in education towards effective personalized learning and enabling educators to adapt to individual differences in SRL during a study episode or educational program.

The results of this study show that self-reports on motivation and strategy use explain 28% of variance in online study behaviour overall versus 18% in academic performance. None of the MSLQ scales on motivation significantly predicted online learning behaviour, whereas several MSLQ scales on strategy use did; namely the scales that measured rehearsal strategies, elaboration strategies, and time and study environment management strategies. Given the prominent role of personal goals, efficacy, and interest in general models for student learning (Fryer, 2017), it is striking that the MSLQ scales on motivation did not explain variance in online trace data. At the same time, these results are in line with previous research on self-reports of motivation in higher education that does show a strong relation with achievement (e.g. Hattie, 2015) but not with study behaviour measured with online trace data (e.g. Cho & Heron, 2015; Zhou & Winne, 2012). This could suggest that the motivational aspects of self-regulatory processes are not captured with online trace data, because of their covert nature. In that respect, it is of importance to recognize the crucial role of self-reports in gaining broad insights in SRL.

At the same time, the results of this study underline the added value of online trace data of learning behaviour. Firstly, online trace data of learning behaviour explained a unique proportion of variance in student academic performance. In this study, the number of logins, number of operations performed on discussion forums and resources accounted for approximately 18% of the variance in academic performance. Although this is substantially less than the 21% found by Gašević and colleagues (2014), it is equal to the amount of variance in academic performance explained by students' self-reports on motivation and strategy use. As expected and along with the findings of the study of Cho & Yoo (2017), where online trace data produced an even more precise prediction of academic performance than students' self-reports, this study provides a strong indication of the potential of online trace data to tap self-regulatory processes, in particular strategy use. The fact that online trace data still predict a unique proportion of variance in academic performance when studied in combination with self-reports might be explained by its potential to bypass a potential upward bias in self-reports of motivation and strategy use due to the overconfidence effect (Taylor & Brown, 1988) as they do not suffer from socially desirable responding bias.

Finally, this study shows that online study behaviour and the relation between self-reports and online trace data varies vastly from week to week. Students' self-reports seem to be mainly based on prior learning experiences within the course, as the reports did explain variation in online learning behaviour prior to the collection of self-reports, but substantially less thereafter. This finding is in line with the expectations around the second research question and fits the theoretical lens provided by Fryer (2017) about the feedback loop between students and their learning experiences. In light of the second central questions of the special issue addressed in this study, we can conclude that the use of self-reports does constrain the analytical choices made with self-report data to some extent: The online trace data revealed large amount of within student variance. Online trace data are therefore an important addition to self-reports in guiding innovations in education towards effective instructional design and personalized learning as they have the potential to give teachers and students real-time insights in strategy use and self-regulation.

An avenue for future research is to combine event-based self-reports and online trace data in measuring SRL. Especially since the results of this study did not identify a relation between online trace data and self-motivational beliefs, potentially due to the weak reliability of some of the MSLQ scales. Furthermore, the identified heterogeneity of the two cohorts in this study warrants further investigation. It is possible that the consent procedure in the first cohort did not yield a representative sample of the students in the first cohort with respect to their propensity to engage in self-regulated learning within the specific context of this course.

Future research is also needed to investigate how real-time measures of SRL through online trace data and self-reports can inform teaching and learning. The LISSA project is a good example where learning analytics were used to support the adviser-student dialogue, in a way that helped motivate students, triggered conversation, and provided tools to add personalization, depth, and nuance to the advising session (Charleer, et al., 2018). Similar efforts based on online trace data or event-based self-reports are scarce, but pivotal to

design and evaluate personalized interventions in online learning environments that promote effective study behaviour.

7. Key points

- Self-report measures of strategy use such as management of time and study environment predict participation in online learning activities, although this relationship is not stable across weeks.
- Student self-reports seem to explain more variance in online learning behaviour of prior weeks than variance in learning behaviour in succeeding weeks.
- Self-report measures and online trace data on self-regulated learning are complementary in predicting study success as they both explain a unique proportion of variance in student academic performance.

8. Acknowledgements

Funding: This work is part of the project 'SURFnet Learning Analytics Hoger Onderwijs', supported by the National Control Unit Educational research (NRO) (project-id 405-17-851).

10. References

Bera, S., & Liu, M. (2006). Cognitive tools, individual differences, and group processing as mediating factors in a hypermedia environment. *Computers in Human Behavior*, *22*, 295-319. http://doi.org/10.1016/j.chb.2004.05.001

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, *54*, 199-231. http://doi.org/10.1016/j.chb.2004.05.001

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). Self-regulation: An introductory overview. In *Handbook of self-regulation* (pp. 1-9). Academic Press.

Charleer, S., Moere, A. V., Klerkx, J., Verbert, K., & De Laet, T. (2017). Learning analytics dashboards to support adviser-student dialogue. *IEEE Transactions on Learning Technologies*, *11*, 389-399. http://doi.org/10.1109/TLT.2017.2720670

Cho, M. H., & Heron, M. L. (2015). Self-regulated learning: the role of motivation, emotion, and use of learning strategies in students' learning experiences in a self-paced online mathematics course. *Distance Education*, *36*, 80-99. http://doi.org/10.1080/01587919.2015.1019963

Cho, M. H., & Yoo, J. S. (2017). Exploring online students' self-regulated learning with self-reported surveys and log files: a data mining approach. *Interactive Learning Environments*, *25*, 970-982. http://doi.org/10.1080/10494820.2016.1232278

Coogan, J., Dancey, C. P., & Attree, E. A. (2005). WebCT: a useful support tool for psychology undergraduates–a Q methodological study. *Psychology Learning and Teaching*, *5*, 61-66. http://doi.org/10.2304/plat.2005.5.1.61

Cornelisz, I., & Van Klaveren, C. (2018). Student engagement with computerized practising: Ability, task value, and difficulty perceptions. *Journal of Computer Assisted Learning*, *34*(6), 828-842. http://doi.org/10.1111/jcal.12292

DeNeui, D. L., & Dodge, T. L. (2006). Asynchronous learning networks and student outcomes: The utility of online learning components in hybrid courses. *Journal of Instructional Psychology*, *33*, 256-260.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*, 391-409. http://doi.org/10.1007/s10648-008-9083-6

Greller, W. & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Journal of Educational Technology & Society, 15*, 42-57. Retrieved from http://www.jstor.org/stable/jeductechsoci.15.3.42

Chauliac, M., Catrysse, L., Gijbels, D., & Donche, V. (2020). It is all in the surv-eye: Can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research, 8*(3), 26–39. http://doi.org/10.14786/flr.v8i3.489

Duncan, T. G., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational psychologist*, *40*, 117-128. http://doi.org/10.1207/s15326985ep4002_6

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*, 6-25. http://doi.org/10.1080/00461520.2011.538645

Ellis, R. A., Marcus, G., & Taylor, R. (2005). Learning through inquiry: student difficulties with online course-based Material. *Journal of Computer Assisted Learning, 21,* 239-252. http://doi.org/10.1111/j.1365-2729.2005.00131.x.

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review*, *16*, 325-345. http://doi.org/10.1007/s10648-004-0003-0

Fryer, L. K. (2017). Building bridges: Seeking structure and direction for higher education motivated learning strategy models. *Educational Psychology Review*, *29*, 325-344. http://doi.org/10.1007/s10648-017-9405-7

Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68-84. http://doi.org/10.1016/j.iheduc.2015.10.002.

Gašević, D., Kovanovic, V., Joksimovic, S., & Siemens, G. (2014). Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative. *The International Review of Research in Open and Distributed Learning*, *15*, 134-176. http://doi.org/10.19173/irrodl.v15i5.1954.

Guerra, J., Hosseini, R., Somyurek, S., & Brusilovsky, P. (2016). An intelligent interface for learning content: Combining an open learner model and social comparison to support self-regulated learning and engagement. In *Proceedings of the 21st international conference on intelligent user interfaces* (pp. 152-163). ACM.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*(2-3), 107-124. http://doi.org/10.1007/s11409-007-9016-7

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Heffner, M., & Cohen, S. H. (2005). Evaluating student use of web-based course material. *Journal of Instructional Psychology*, *32*, 74-82.

Hilpert, J. C., Stempien, J., van der Hoeven Kraft, K. J., & Husman, J. (2013). Evidence for the latent factor structure of the MSLQ: A new conceptualization of an established questionnaire. *SAGE Open*, *3*, 2158244013510305.

Hoskins, S. L., & Van Hooff, J. C. (2005). Motivation and ability: which students use online learning and what influence does it have on their achievement? *British Journal of Educational Technology*, *36*, 177-192. http://doi.org/10.1111/j.1467-8535.2005.00451.x.

Howard-Rose, D., & Winne, P. H. (1993). Measuring component and sets of cognitive processes in self-regulated learning. *Journal of Educational Psychology*, *85*, 591. http://doi.org/10.1111/j.1464-0597.2005.00205.x

Huon, G., Spehar, B., Adam, P., & Rifkin, W. (2007). Resource use and academic performance among first year psychology students. *Higher Education*, *53*, 1-27. http://doi.org/10.1007/s10734-005-1727-6

Iaconelli, R., & Wolters, C. A. (2020). Insufficient effort responding in surveys assessing self-regulated learning: Nuisance or fatal flaw? *Frontline Learning Research, 8*(3), 105–127. http://doi.org/10.14786/flr.v8i3.521

Itmazi, J. A., & Megías, M. G. (2005). Survey: Comparison and evaluation studies of learning content management systems. *Unpublished manuscript*.

Jamieson-Noel, D., & Winne, P. H. (2003). Comparing Self-Reports to Traces of Studying Behavior as Representations of Students' Studying and Achievement. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology.* http://doi.org/10.1024//1010-0652.17.34.159

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough. Pitfalls of learning analytics dashboards in the educational practice. In É. L., H. D., K. V., J. B., & M. P-S. (Eds.), *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings* (Lecture Notes in Computer Science (LNCS); Vol. 10474). Cham: Springer International Publishing AG. http://doi.org/10.1007/978-3-319-66610-5_7

Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018). License to evaluate: preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 31-40). ACM. http://doi.org/10.1145/3170358.3170421

Jiang, L., Elen, J., & Clarebout, G. (2009). The relationships between learner variables, tool-usage behaviour and performance. *Computers in Human Behavior*, *25*, 501-509. http://doi.org/10.1016/j.chb.2008.12.010

Jupp, V. (2006). *The Sage dictionary of social research methods.* Sage.

Kim, D., Yoon, M., Jo, I. H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers & Education*, *127*, 233-251. http://doi.org/10.1016/j.compedu.2018.08.023.

Knight, J. (2010). Distinguishing the learning approaches adopted by undergraduates in their use of online resources. *Active Learning in Higher Education, 11,* 67–76. http://doi.org/10.1177/1469787409355873

Lust, G., Collazo, N. A. J., Elen, J., & Clarebout, G. (2012). Content management systems: enriched learning opportunities for all? *Computers in Human Behavior, 28,* 795-808. http://doi.org/10.1016/j.chb.2011.12.009

Lust, G., Elen, J., & Clarebout, G. (2013). Regulation of tool-use within a blended course: Student differences and performance effects. *Computers & Education, 60,* 385-395. http://doi.org/10.1016/j.compedu.2012.09.001

McKeachie, W. J., Pintrich, P. R., & Lin, Y. G. (1985). Teaching learning strategies. *Educational Psychologist, 20,* 153-160. http://doi.org/10.1207/s15326985ep2003_5

Molenda, M. (2008). Historical foundations. In M. J. Spector, M. D. Merrill, J. van Merrienboer, & M. P. Driscoll (Eds.). *Handbook of research for educational communications and technology* (pp. 5–20). Routledge.

Moeller, J., Dietrich, J., Viljaranta, J., & Kracke, B. (2020). Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Frontline Learning Research, 8*(3), 63–85. http://doi.org/10.14786/flr.v8i3.529

Muis, K. R., Winne, P. H., & Jamieson-Noel, D. (2007). Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *British Journal of Educational Psychology, 77,* 177–195. http://doi.org/10.1348/000709905X90876

Munoz, KD, & Van Duzer, J. (2005). *BlackBoard vs. Moodle: A Comparison of Satisfaction with Online Teaching and Learning Tools.* Humboldt State University.

Panadero, E. (2017). A review of self-regulated learning: six models and four directions for research. *Frontiers in Psychology*, *8*, 422. http://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Klug, J., & Järvelä, S. (2016). Third wave of measurement in the self-regulated learning field: when measurement and intervention come hand in hand. *Scandinavian Journal of Educational Research, 60,* 723-735. http://doi.org/10.1080/00313831.2015.1066436

Pintrich, P. R. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*, 544. http://doi.org/10.1037/0022-0663.92.3.544

Pintrich, P. R., McKeachie, W. J., & Lin, Y. G. (1987). Teaching a course in learning to learn. *Teaching of Psychology*, *14*, 81-86. http://doi.org/10.1207/s15328023top1402_3

Rogiers, A.; Merchie, E., & van Keer, H. (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research, 8*(3), 40–62. http://doi.org/10.14786/flr.v8i3.527

Schmitz, M., Scheffel, M., van Limbeek, E., van Halem, N., Cornelisz, I., van Klaveren, C., ... & Drachsler, H. (2018). Investigating the Relationships Between Online Activity, Learning Strategies and Grades to Create Learning Analytics-Supported Learning Designs. In *European Conference on Technology Enhanced Learning* (pp. 311-325). Springer, Cham.

Taylor, S. E. & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103,* 193–210. http://doi.org/10.1037/0033-2909.103.

Tock, J. L., & Moxley, J. H. (2017). A comprehensive reanalysis of the metacognitive self-regulation scale from the MSLQ. *Metacognition and Learning*, *12*, 79-111. http://doi.org/10.1007/s11409-016-9161-y

Veenman, M. (2011). Learning to self-monitor and self-regulate. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction (pp. 197–218).* New York: Routledge.

Vriesema, C. C., & McCaslin, M. (2020) Experience and meaning in small-group contexts: Fusing observational and self-report data to capture self and other dynamics. *Frontline Learning Research, 8*(3), 128–141. http://doi.org/10.14786/flr.v8i3.493

Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in psychology web-based classes. *Journal of Educational Psychology*, *92*, 137. http://doi.org/10.1037/0022-0663.92.1.137.

Winne, P. H., & Baker, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining*, *5*, 1-8. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/28

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation.* Academic Press.

Zhou, M., & Winne, P. H. (2012). Modelling academic performance by self-reported versus traced goal orientation. *Learning and Instruction, 22,* 413-419. http://doi.org//j.learninstruc.2012.03.004

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45,* 166–183. http://doi.org/10.3102/0002831207312909

APPENDIX A

Cronbach's Alpha per subscale of the MSLQ per cohort

		Cohort 2016		Cohort 2017	
Subscale	**Definition (adapted from Duncan & McKeachie, 2015)**	N	α	N	α
Intrinsic Goal Orientation	Goal orientation refers to the student's perception of the reasons why she is engaging in the course. Having intrinsic goal orientation means participation is an end all to itself, rather than participation being a means to an end. Sample item: In a class like this, I prefer course material that really challenges me so I can learn new things.	341	.65	466	.58
Extrinsic Goal Orientation	Extrinsic goal orientation complements intrinsic goal orientation, with the degree to which the student perceives herself to be participating in a task for reasons such as grades, rewards, performance, evaluation by others, and competition; meaning engaging in the course is the means to an end. Sample item: Getting a good grade in this class is the most satisfying thing for me right now.	347	.73	465	.59
Task Value	Task value differs from goal orientation in that task value refers to the student's evaluation of the how interesting, how important, and how useful the course is. Sample item: I think I will be able to use what I learn in this course in other courses.	342	.81	458	.79
Self-Efficacy	Self-efficacy is a self-appraisal of one's ability to master a task. Self-efficacy includes judgments about one's ability to accomplish a task as well as one's confidence in one's skills to perform that task. Sample item: I believe I will receive an excellent grade in this class.	341	.86	462	.92
Test Anxiety	Test anxiety refers to students' negative thoughts that disrupt performance, and the affective and physiological arousal aspects of anxiety. Sample item: When I take a test, I think about how poorly I am doing compared with other students.	345	.82	464	.82
Control of Learning Beliefs	Control of learning concerns the belief that outcomes are contingent on one's own effort, in contrast to external factors such as the teacher. If students believe that their efforts to study make a difference in their learning, they should be more likely to study more strategically and effectively. Sample item: If I study in appropriate ways, then I will be able to learn the material in this course.	345	.65	470	.69
Metacognitive Self-Regulation	Metacognitive self-regulation refers to planning, monitoring, and regulating. Planning activities such as goal setting and task analysis help to activate, or prime, relevant aspects of prior knowledge that make organizing and comprehending the material easier.	331	.71	375	.74

	Monitoring activities include tracking of one's attention as one read, and self-testing and questioning: these assist the learner in understanding the material and integrating it with prior knowledge. Regulating refers to the fine-tuning and continuous adjustment of one's cognitive activities. Sample item: When reading for this course, I make up questions to help focus my reading.				
Rehearsal	Basic rehearsal strategies involve reciting or naming items from a list to be learned. These strategies are best used for simple tasks and activation of information in working memory rather than acquisition of new information in long-term memory. Sample item: When I study for this class, I practice saying the material to myself over and over.	338	.65	377	.65
Elaboration	Elaboration strategies help students store information into long-term memory by building internal connections between items to be learned. Elaboration strategies include paraphrasing, summarizing, creating analogies, and generative note taking. Sample item: When I study for this class, I pull together information from different sources, such as lectures, readings, and discussions.	335	.66	371	.68
Organization	Organization strategies help the learner select appropriate information and also construct connections among the information to be learned. Examples of organizing strategies are clustering, outlining, and selecting the main idea in reading passages. Organizing is an active, effortful endeavour, and results in the learner being closely involved in the task. Sample item: When I study the readings for this course, I outline the material to help me organize my thoughts.	339	.64	381	.66
Critical Thinking	Critical thinking refers to the degree to which students report applying previous knowledge to new situations in order to solve problems, reach decisions, or make critical evaluations with respect to standards of excellence. Sample item: I often find myself questioning things I hear or read in this course to decide if I find them convincing.	337	.77	380	.79
Time and Study Environment	Time management involves scheduling, planning, and managing one's study time. This includes not only setting aside blocks of time to study, but the effective use of that study time, and setting realistic goals. Time management varies in level, from an evening of studying to weekly and monthly scheduling. Study environment management refers to the setting where the student does her class work. Sample item: I usually study in a place where I can concentrate on my course work.	341	.85	375	.79
Effort Regulation	Effort regulation is self-management in the face of distractions and uninteresting tasks and reflects a commitment to completing one's study goals.	341	.70	375	.73

	Sample item: I work hard to do well in this class even if I don't like what we are doing.				
Peer Learning	Peer learning involves peer collaboration and dialogue, which can help a learner reach insight one may not have attained on one's own. Sample item: When studying for this course, I often try to explain the material to a classmate or a friend.	339	.66	380	.68
Help Seeking	Help seeking involves knowing when one doesn't know something and being able to identify someone to provide them with some assistance. Sample item: I ask the instructor to clarify concepts I don't understand well.	338	.71	378	.63

APPENDIX B

Example questions from the 2016-2017 exam

An example of a knowledge question:*
Fill in the blanks: The ….I…. test is used for testing the difference in proportions between dependent samples and the ….II…. test for testing the difference in proportions between small independent samples.

a. I: Binomial, II: Chi-squared
b. I: Chi-squared, II: binomial
c. I: McNemar, II: Fisher's exact
d. I: Fisher's exact, II: McNemar

An example of an insight question:
In order to evaluate differences in study success across different programs offered at the VU, an equal number of students are randomly selected from each program and asked to participate in the research. This is an example of:
a. Systematic random sampling
b. Cluster sampling
c. Stratified random sampling
d. Multi-stage sampling

An example of a calculation question:
Research results have revealed that intelligence in the Netherlands is normally distributed. The mean IQ score is 100 with a standard deviation of 15. John has an IQ-score of 122.5. What percentage of people in the Netherlands will have an IQ lower than that of John?
a. 3.34 %
b. 6.68 %
c. 93.32 %
d. 96.66 %

*The correct answer is provided in bold.

Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 8 No. 3 special issue (2020) 164 - 173
ISSN 2295-3159

Commentary: A Proposed Remedy for Grievances about Self-Report Methodologies

Philip H. Winne[1]

[1]Simon Fraser University, Canada

Abstract

This special issue's editors invited discussion of three broad questions. Slightly rephrased, they are: How well do self-report data represent theoretical constructs? How should analyses of data be conditioned by properties of self report data? In what ways do interpretations of self-report data shape interpretations of a study's findings? To approach these issues, I first recap the kinds of self-report data gathered by researchers reporting in this special issue. With that background, I take up a fundamental question. What are self-report data? I foreshadow later critical analysis by listing facets I observe in operational definitions of self-report data: nature of the datum, topic, property, setting or context, response scale, and assumptions setting a stage for analyzing data. Discussion of these issues leads to a proposal that ameliorates some of them: Help respondents become better at self reporting.

Keywords: self-report data; Likert scale; think aloud protocol

Corresponding author:Professor Phil Winne Simon Fraser University, EDB 8515, 8888 University Drive, Burnaby, BC V5A 1S6 Canada email: winne@sfu.ca *DOI: 10.14786/flr.v8i3.631*

1. The Landscape of Self-Reports Represented in this Special Issue

The most common forms of self-report data are surveys and think-aloud protocols. The former were popular among articles in this special issue. Only one study used think aloud procedures.

Chauliac, et al. (2020) asked participants to rate how frequently they applied various cognitive processes while studying textual information. Durik and Jenkins (2020) administered survey items calling for respondents to the degree to which they agreed with statements describing interest in several areas of study: astronomy, biology, math and psychology; and a single item inviting respondents to classify how certain they were about the set of ratings of interest. Fryer and Nakao (2020) investigated several judgments students made about a course. Topics ranged over interest, meaningfulness and other dimensions. Their focus was on possible differences arising due to different quantitatively described response formats – a conventional labeled categorical (Likert) scale, slider, swipe and visual analog. Iaconelli and Wolters (2020) administered three surveys inquiring about respondents' ratings of agreement and confidence about motivational constructs and self-regulated learning. Moeller, et al., (2020) gathered participants' ratings of the degree to which statements about interest in, emotional investment in, and value of a course applied to them. Rogiers, et al., (2020) administered a survey instrument on which participants rated how much they agreed with statements describing their use of several cognitive and metacognitive processes. These researchers also gathered participant's think aloud accounts about how participants studied an informative text. van Halem, et al., (2020) administered the well-known Motivated Strategies for Learning Questionnaire which includes various subscales of Likert items describing constructs within the arenas of motivation, cognition and metacognition. Participants in Vriesema and McCaslin's (2020) study responded to a survey including Likert response items about anxiety and selected from among a set of 20 sentences ones that described perceptions about participation in a small group activity.

Various features can more thoroughly discriminate the nature of self-reporting as a process and data these studies represent. In the next section, I propose a typology for these and other self-report methodologies.

2. Facets of a Self-Report Datum

Among facets of self-report data, primus inter pares (first among equals) is reliance on language. A self-report datum is a participant's verbal utterance (think aloud) or recorded response to a spoken (interview) or written (survey, diary, experience sampling) invitation to describe a state or an event. The invitation to self report and the report itself are couched in language. No psychometric computation can remedy or precisely quantify indefiniteness arising from the dependence of self-report data on inherent elasticity and nuance of natural language. Consequently, validity of interpretations grounded in self-report data is lessened in proportion to this source of unreliability. A variety of measures reported in this special issue and elsewhere in the literature that researchers intend to parallel or contrast to self-report data are not self-report measures according to this conceptualization. Examples include: electrodermal signals, heart rate, various electromagnetic records of brain activity and data derived from tracking eye gear.

Topics described by self-report data range very widely. Two main divisions are apparent: states and events. States may be internal, for example, a mood or physical condition. Vriesema and McCaslin's (2020) participants made a dichotomous decision whether their "stomach felt funny" or "head hurt" during a group activity. States also can be external, such as a characteristic of a learning situation or the availability of needed information. Iaconelli and Wolter's (2020) participants rated desire for more time to complete schoolwork and finishing assignments right before deadlines. Events are marked by changes internal to the respondent (increased anxiety, decreased certainty the effectiveness of a studying tactic) and in the

environment (reviewing previously studied content, searching for information using an online search engine). Students in Rogiers, et al., (2020) study described marking important words and rehearsing information. To the degree that language is indefinite, so, too, are self-report data. Included within articles published in this special issue are: affective states and emotional reactions (e.g., interest, enjoyment, worry, pride), physiological states (stomach upset), behavioral events (summarizing content, copying notes, engagement with group members), and cognitive and metacognitive events (rehearsing content, judging extent and qualities of learning, planning). Other studies range further afield. Variance in the meaning of these features almost surely differs among participants within a study and across contexts. What kinds of tasks are considered schoolwork and which are not? What time interval between finishing an assignment and a deadline is "right before"? What makes words important? How do individuals' constructions of these concepts differ in ways that matter to the research questions each study investigated?

Properties of topics respondents self report about also vary. Included in the research reported in this special issue are: frequency of an event, intensity of states, certainty of knowledge or about a future state, fit to one's view of self, typicality of the event or state, and appropriateness of some behavior or feeling relative to a setting. For example, Moeller et al., (2020) experience sampling probes asked respondents to rate their understanding of, liking for, effort put toward learning, annoyance at learning, emotional cost to learn and future value of learning particular subject matter content. The wider literature adds to liberally to these topics.

Invitations to self report refer to a setting or context relative to which the report is forged. Settings or contexts range over two dimensions. One dimension is whether a specific setting frames the self-report. Examples are an experience in the immediate past, such as a just-completed session of collaborative work or a class period that has just finished. Moeller and colleagues (2020) minimized the time interval with their experience sampling method, as did Chauliac et al. (2020) by administering their survey after participants had just completed a studying task. The other end of this spectrum is a generalized setting, such as studying or life in school. Some items in the Motivated Strategies for Learning Questionniare used in van Halem et al. (2020) study ask respondents to consider "a typical course." Durik and Jenkins (2020) asked students to rate how strongly they had "always been fascinated with mathematics" and how much they were "really looking forward to learning more about mathematics." The second dimension of the setting or context is whether that setting or context is one the respondent has personally experienced as opposed to one the respondent is asked to imagine. A prime example of a personally experienced context is the popular think aloud protocol like that used in the study by Rogiers et al. (2020). Participants work on a task and talk about what they think or do while they engage with it or, sometimes, retrospectively, quite soon after the participant disengages from the task. Two variants of this latter case vary the delay between task completion and self report. Under the methodology of experience sampling, like that in Moeller et al.'s (2020) study, learners are notified at random points in time, commonly by a "beeper" or mobile phone notification, to write out or make an audio recording about an experience just completed. Under a diary methodology, learners self report at periodic intervals, such as on the weekend. When respondents are asked to imagine a setting or context, they predict what would be the case if the experience actually happened.

Various response scales are used as metrics for self-report data. Fryer and Nakao's (2020) study illustrates a direct investigation of this. In cases where respondents;' utterances are analyzed by researchers who do not adopt an a priori classification, the researcher invents a categorical metric based on responses generated by one or multiple respondents. Categorical bins of data are sometimes described as themes. In this case, respondents do not know when they respond how their self reports will be binned, so responses can not be biased by a response format. In other cases, respondents are fully aware how their self reports are "scored" because the response scale is used to provide a response. Some scales call for selecting one option from a set of categories; sex, academic major and race are examples. Other scales are ordinal, such as the extensively used Likert scale. Here, the response is expressed as a relative quantity, e.g., strongly agree/disagree, rarely/almost always, or very unlike/like me. When response scales are ordinal, an important decision the researcher makes is whether to permit the respondent to declare neutrality or indifference by

offering an odd number of ratings versus an even number of ratings which typically precludes that option. Some researchers ask for actual counts of events. I believe respondents cannot be accurate in this case unless their memory is perfect, a rare if not implausible quality.

3. Assumptions, Issues and Complaints Regarding Self-Report Data

I stipulate for sake of argument that respondents respond to invitations to self report to the best of their abilities. To borrow a common phrase from television shows about witnesses in American court cases, when respondents self report, they intend to "tell the truth, the whole truth and nothing but the truth."

Likert scaled data almost always are analyzed using conventional arithmetic operations, such as forming a subscale by summing or averaging responses to several items. This was true of studies reported in this special issue. Arithmetic operations require data have properties of a continuous interval scale. That is, units are differentiable (e.g., on a 100-point scale, 67 is different from 68) and differences between any adjacent pair of scale points is assumed to measure the same amount (e.g., the same difference is spanned between 67 and 68 and between 97 and 98). Researchers usually sidestep this issue by reasoning this way – adding a large number of ordinal responses across separate items, say ratings of 1 to 5 over 10 items, elongates the scale so that it approximates a continuous scale with intervals of equal size. In my example, summing over 10 items increases the maximum scale length to 50 which affords a sufficient approximation to properties needed to use conventional arithmetic. This sleight of hand requires all the items represent one underlying dimension. To avoid the "apples and oranges" problem, the scale must be unidimensional to warrant adding responses. Researchers attempt to test this requirement by computing a measure of internal consistency reliability or dimensionality. Common approaches include computing Cronbach's alpha coefficient or doing a principal components or factor analysis. However, such computations put the cart before the horse. The assumption to be tested must hold to do these calculations because they require applying arithmetic operations to response values. The ploy is therefore tautological. As well, it often suffers important flaws (see Liddell & Kruschke, 2018).

When respondents are invited to think aloud, write diary entries or otherwise generate free responses, all but a tiny fraction of studies bin instances by aggregating across respondents. Rogiers et al. created 11 bins to differentiate text-learning strategies. Implicit in this practice is a critical assumption, namely, variance across respondents and points-in-time/within-task are unimportant. This is analogous to the issue just discussed about Likert data. As well, a tacit assumption rarely explicitly addressed by the researcher is there are no interactions among facets. That is, self reports in one bin do not correlate with others in a different bin and the kind of self report placed in one bin is not conditioned on another bin. For example, solicitations are independent of replies, recognition of an obstacle has no relation to solutions that overcome obstacles or admissions that obstacles cannot be overcome. My sample is small but I have never observed researchers test both assumptions.

Other assumptions, usually unstated and commonly untested, underlie researchers' interpretations about bins of utterances or factors/components generated by a quantitative method. An utterance or survey item is assigned to one and only one bin, subscale or factor/component when the researcher's analysis signals it can belong to multiple bins (Fryer & Nakao, 2020; Rogiers et al., 2020). This practice simplifies analyses but may well oversimplify what respondents mean. When self reports that could be classified into multiple bins/components/factors are excluded from some of those containers and assigned to a single bin, respondents' data are biased by that operational definition.

All self report data become available because a researcher invites respondents to report. Before issuing the invitation, the researcher cannot know whether information encapsulated in the requested report would have existed absent the researcher's invitation. This raises a conundrum. Some self reports may be

outright fabrications. On being asked to report about something that was not in a respondent's working memory or awareness, respondents may reply only to fulfill a social demand created by the researcher's request. If this is the case, instrumentation creates a state or event that is reported rather than externalizing a report about a state or event that existed before the respondent was invited to describe it. Consider this item from Rogiers et al.'s (2020) scale about metacognitive monitoring: "I managed to learn the text in a good way." The time point for a respondent to make this judgment is unspecified so a learner may honestly respond about monitoring during study or make the judgment when the survey is administered. Valid interpretations about how a learner processes information will be challenged.

This possibility raises a perplexing issue particularly in the context of agentic behavior like self-regulated learning (SRL; Winne, 2018). An agent's behavior or experience can be altered as the agent becomes aware of characteristics of that behavior and experience. So, would a learner have considered the topic and properties of an experience or event if s/he had not been asked? In what cases and to what degrees are self reports possibly epiphenomenal?

Regarding think aloud reports, Ericsson and Simon (1993; see also Fox, et al., 2011) were acutely aware of the foregoing possibility. After scrutinizing research available at the time, they concluded: "With great consistency, this evidence demonstrates that verbal data are not in the least epiphenomenal but instead are highly pertinent to and informative about subjects" cognitive processes and memory structures" (p. 220). But they also note an important caveat.

> When subjects verbalize directly only the thoughts entering their attention as part of performing the task, the sequence of thoughts is not changed by the added instruction to think aloud. However, if subjects are also instructed to describe or explain their thoughts, additional thoughts and information have to be accessed to produce these auxiliary descriptions and explanations. As a result, the sequence of thoughts is changed, because the subjects must attend to information not normally needed to perform the task (p. xiii).

Rogiers et al. (2020) wisely engaged participants in a practice session to clarify the process of thinking aloud. They also adopted the common practice of prompting participants to "'verbalize everything that you are doing or thinking' or 'keep thinking aloud'" when the researcher perceived there were "(a) meaningful silences or (b) certain nonverbal behaviours took place (i.e., frowning, repeatedly turning the text page, staring)." It might be argued this challenges Ericsson and Simon's caveat. Translating a state or event that has non-linguistic form into language involves considering what words are best to use. It is unclear whether words validly represent the learner's experience and whether monitoring the qualities of that translation adds information into the cognitive arena not present before the learner was prompted to think aloud.

Every paper-and-pencil or computer-delivered questionnaire item asks respondents to describe properties of a thought, such as its generality, frequency or intensity. Sometimes, respondents unintentionally make up answers. A relevant case arose in a study comparing self reports to logs of online behavior (Winne & Jamieson-Noel, 1982). In this study, learners using software to study could view objectives for learning by clicking a button. The software logged this action if the learner clicked the button. After studying and taking an achievement test, we asked learners how helpful they found the objectives as guides to learning. Several participants responded the objectives were helpful but the log of their data showed they never accessed the objectives. A more recent study reported less blatant but still important differences between online (logged) behaviors and self reports about those behaviors: "Self-reports on prospective questionnaires show poor across method convergence with on-line thinking-aloud and observational data, obtained from students solving mathematics problems. These results are in line with earlier multi-method studies for reading and mathematics (see above). Likewise, self-reports on the retrospective questionnaire do not converge with observational data" (Veenman & van Cleef, 2019, p. 698).

A further complication arises when researchers gather self-report data to characterize SRL. I model an elemental SRL event as an IF-THEN production. Conditions (IFs) are the context in which a learner applies a particular cognitive operation to particular information (THEN). What was just addressed relates to THENs, an action learners perform. On the other side of this model, every self report methodology specifies conditions, IFs, the context within which the learner is to reply. As noted earlier, these may be set out for the learner in general terms, e.g., "this course" as in the Motivated Strategies for Learning Questionnaire (see van Halem et al., 2020) or "the lecture contents of the past couple of minutes" (Moeller et al., 2020, p. 6). When a learner responds, it is reasonable to infer some particular features of a setting, IFs, influence the learner's response. What are those features? Is it reasonable to aggregate data across learners if there is variance in those features across learners? When the context is described for respondents as "this course," to all learners characterize "the" course in sufficiently similar ways to warrant treating responses as if the conditions are the same? Researchers lack data about what specifically each learner construes as IFs when responding to most survey items. Assuming those conditions are the same when respondents have the same response, identical THENS, is an instance of a logical fallacy, post hoc ergo prompter hoc. The same fallacy applies when researchers compute stability coefficients (test-retest) to characterize reliability of self-report data.

A great number of studies using self-report data correlate those data with other, usually, outcome variables such as achievement or satisfaction with a long-term prior experience. In the studies published in this special issue and throughout the literature, language commonly used to express such results casts self report data as "accounting for" or "explaining" variance in another variable. These and analogous phrases implicitly but invalidly refer to the construct represented by self-report data as a cause. Misleading phrasing about correlational findings that invites understanding them as causal is a common gaffe (Robinson et al., 2007).

This matter becomes even more muddled when statistical features of self-report data are not recognized. First, like all other data, self-report inherently have noise. Some variance arises due to randomness and this attenuates the magnitude of relations. It might be suggested this challenge could be met by correcting for attenuation, but that operation actually muddles interpretation (see Winne & Belfry, 1982). Second, it is often the case that self-report data are among other predictors used to predict an outcome. Multiple regression and similar models are common statistical methods applied in this case, as was true for studies in this special issue (Chauliac et al., 2020; Durik & Jenkins, 2020; van Halem et al., 2020; Moeller et al., 2020; Vriesema & McCaslin, 2020). In these types of analysis, each predictor is residualized for every other predictor. Interpretations of results of this analysis almost always fail to acknowledge the statistical output describes a residualized variable, not the original (Winne, 1983). Accurate phrasing relating to a beta coefficient such as, "Self-reported elaboration residualized for the self-reported importance of other studying methods, sex of respondent and an indication of academic ability" appears in print as "elaboration … ." Validity is even more strained by this oversight because it is not explicit to readers that the relation concerns a mathematically residualized construct that is not the same as the original construct.

4. Conclusions

The editors asked: How well do self-report data represent theoretical constructs? How should analyses of data be conditioned by properties of self-report data? In what ways do interpretations of self-report data shape interpretations of a study's findings?

"Construct" is the pivotal word in the opening question of this trio. Data, whether self reported and otherwise, are realized through operations a researcher designs and the correspondence between that operational definition and its implementation that realizes data.

As implied by the title of Gitelman's (2013) edited volume, "Raw Data" Is an Oxymoron, data are not raw in the sense of lacking bias. Theory is the muse that inspires gathering particular data in particular ways. From this perspective, all data inherently have some bias because they originate in a particular theory that conceptualizes constructs and characterizes forms for representing them as data.

Features of the physical, mathematical and psychological realms shape how researchers can obtain data sought to investigate a theory in ways shaped by multiple theories. The veracity with which realized self report data represent a particular construct depends on instrumentation writ large – instructions about how and when to respond, the setting in which a respondent reports, response format, and other values for facets described previously. This concern with generalizability was a core message of Cronbach et al. (1972) views of generalizability, their extension to classical test theory's notion of reliability. Self-report data (as well as other forms of data) should be addressed not from a perspective of "the" reliability but a recognition of need to investigate generalizability. The fundamental question is: Which facets and over what range of a facet's values is unwanted (or uninterpretable) variance introduced into data (see Winne, 2018)?

In the context of self-report data, questions about facets and their relations to generalizability arise quickly. What is judged an "authentic" setting and what is not? Who decides? Do instructions cause a respondent to report memories as best they can be recalled or invent a false but plausible "memory" that fits the invitation to self report? Was the protocol designed to invite a self report realized as designed (sometimes labelled fidelity of treatment implementation)? What facets of an operational design generate or suppress variance in the self-report data?

Replies to questions about how well data represent constructs will not be simple. While it is taxing, researchers should test facets of self-report data within their research. Triangulation with non-self-report data is one approach, as illustrated in several articles appearing in this special issue. But triangulation is a hedge against issues of generalizability rather than an escape from them. Identifying relevant facets is a precursor to improving opportunities to validly interpret data and analyses of data.

Analyses of self-report data can be improved. A key is to consider which laws of mathematics or, for categorical data, logic apply to self-report data as they are operationalized. The time to raise this question when designing research rather than after data have been gathered. Researchers should more explicitly realize and describe in their publications the calculus applied to self-report data. Mathematical manipulations of data are part of the data's operational definition. This aspect of operational definitions have bearing on opportunities to validly interpret results of calculations. I illustrated this for cases where self-report data join other predictors in a multiple regression model. Data residualized in such models are transformations of "data-at-the-first-instance" (a cumbersome label I hope may spark a good replacement for the more common label "raw data"). It will be helpful to readers if researchers remind them all components of operational definitions.

An addendum to this recommendation is to entertain, when operational definitions of self-report data are being engineered, alternative analytical methods that might be applied to self-report data. For example, with Likert-scaled data from surveys, what are trade-offs if items are examined using a principal components analysis versus a cluster analysis?

Two further considerations can be raised when analyzing self-report data in verbal forms, such as transcripts of think aloud data and replies to interview questions that are freely structured by the respondent. Commonly, researchers labor to identify theoretically sensible bins or themes, then investigate interrater agreement before dividing up work on the corpus. At both stages, some, often large portions of data are discarded because those reports don't fit bins. Respondents, however, believed their accounts were relevant to the researcher's invitation to report. Disregarding these data represents another form of bias and this practice may mask or misconstrue what respondents deemed representative of their theory.

Second, bins of self-report data sometimes disregard temporal development and contingencies among bins. This may be particularly important in think-aloud data where unfolding episodic content is central to the respondent's experience. Self-report data should be investigated for contingencies and trajectory beyond statically binning segmented self-reports.

The final question the editors posed asked about ways interpretations of self-report data shape interpretations of a study's findings. At first blush, this might seem trivial if the word "findings" is taken to mean what appears in the discussion section of an article or chapter. Those "findings" are interpretations of analyses of self-report (and other) data, so interpretations of self-report data are directly related to the study's findings. A different perspective reflects what I observe more frequently in the research literature.

A great deal of research that analyzes and interprets self-report data is carried out to investigate a research hypothesis. Hypotheses are shaped in the first place by a theory the researcher chooses and uses to guide the investigation. As previously discussed, theoretical lenses shape decisions about what data merit collecting in the first place, instrumentation used to gather those data and warrants for features of analyses of data. It is important to keep in mind that theory sharpens some phenomena, blurs others and renders the rest invisible by classifying them as unimportant. My answer to the editors' final question is that a study's findings inevitably and substantially are shaped by a researcher's interpretations about what self-report data are worth collecting. In other words, findings in the past shape theories that shape interpretations in the future.

5. Coda

Science may someday develop instruments that accurately "read" human brain activity in a way that can reveal exactly what a person is thinking. Current instruments – face readers, gaze trackers, and other physiological sensors – in my judgment, are not capable of that task. For the time being, learning science must rely partly on what people tell about their thoughts and feelings.

When researchers collect self-report data, they depend on the respondent to "know thyself." In less quaint terms, the respondent is a critical cog in a system that generates self-report data. I forecast learning science can better understand self report data and more prudently use them by seeking a fuller account about why knowing one's self is difficult, and how people can more fully and more accurately come to know themselves. It follows that one approach to remedying some grievances I presented is to investigate how to help respondents – the key component within a system of instrumentation that develops self-report data – improve self reporting.

Acknowledgments

Foundations for this article were developed with financial support provided over many years by the Social Sciences and Humanities Council of Canada and Simon Fraser University.

References

Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment*, *21*(1), 19-33. https://doi.org/10.1080/10627197.2015.1127751

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Chauliac, M., Catrysse, L., Gijbels, D., & Donche V. (2020). It is all in the surv-eye: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*, *8*(3), 26 – 39. https://doi.org/10.14786/flr.v8i3.489

Durik, A. M., & Jenkins J. S. (2020). Variability in Certainty of Self-Reported Interest: Implications for Theory and Research. *Frontline Learning Research*, *8*(3) 85-103. https://doi.org/10.14786/flr.v8i3.491

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised edition). MIT Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*, 316-44. https://doi.org/10.1037/a0021663

Fryer, L. K., & Nakao K. (2020). The Future of Survey Self-report: An experiment contrasting Likert, VAS, Slide, and Swipe touch interfaces. *Frontline Learning Research*, 8(3),10-25. https://doi.org/10.14786/flr.v8i3.501

Gitelman, L. (Ed.). (2013). *"Raw Data" Is an Oxymoron.* The MIT Press: Cambridge, MA. https://doi.org/10.7551/mitpress/9302.001.0001

Van Halem, N., van Klaveren, C., Drachsler H., Schmitz, M., & Cornelisz, I. (2020). Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data. *Frontline Learning Research*, *8*(3), 140-163. https://doi.org/10.14786/flr.v8i3.497

Iaconelli, R., & Wolters C.A. (2020). Insufficient Effort Responding in Surveys Assessing Self-Regulated Learning: Nuisance or Fatal Flaw? *Frontline Learning Research*, 8(3), *104 – 125.* https://doi.org/10.14786/flr.v8i3.521

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., , De Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*, 139–151. https://doi.org/10.1080/00461520701416231

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328-348. https://doi.org/10.1016/j.jesp.2018.08.009

Moeller, J., Viljaranta, J., Kracke, B., & Dietrich, J. (2020). Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Frontline Learning Research*, *8*(3), 63-84. https://doi.org/10.14786/flr.v8i3.529

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. (2007). The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal*, *44*(2), 400-413 https://doi.org/10.3102/0002831207302174

Rogiers, A., Merchie, E., & Van Keer H. (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research*, *8*(3), 40 – 62. https://doi.org/10.14786/flr.v8i3.527

Veenman, M. V. J., & van Cleef, D. (2018). Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM*, *51*(4), 691-701. https://doi.org/10.1007/s11858-018-1006-5

Vriesema, C.C., & McCaslin, M. (2020) Experience and Meaning in Small-Group Contexts: Fusing Observational and Self-Report Data to Capture Self and Other Dynamics. *Frontline Learning Research*, *8*(3), 126-139. https://doi.org/10.14786/flr.v8i3.493

Winne, P. H. (1983). Distortions of construct validity in multiple regression analysis. *Canadian Journal of Behavioural Science*, *15*, 187-202. https://doi.org/10.1037/h0080736

Winne, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, *19*, 125-134. https://www.jstor.org/stable/1434905?seq=1#metadata_info_tab_contents

Winne, P. H., & Jamieson-Noel, D. L. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology*, *27*, 551-572. https://doi.org/10.1016/S0361-476X(02)00006-1

Winne, P. H. (2018). Paradigmatic issues in state-of-the-art research using process data. *Frontline Learning Research*, *6*, 250-258. https://doi.org/10.14786/flr.v6i3.551

Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol. 8 No. 3 special issue (2020) 174 - 184
ISSN 2295-3159

Commentary: Measurement and the Study of Motivation and Strategy Use: Determining If and When Self-report Measures are Appropriate

Peggy N. Van Meter[a]

[a]Pennsylvania State University, USA

Abstract

The goal of this special issue is to examine the use of self-report measures in the study of motivation and strategy use. This commentary reviews the articles contained in this special issue to address the primary objective of determining if and when self-report measures contribute to understanding these major constructs involved in self-regulated learning. Guided by three central questions, this review highlights some of the major, emergent themes regarding the use of self-report. The issues addressed include attention to evidence for construct validity, the need to consider broad methodological factors in the collection and interpretation of self-report data, and the innovations made possible by modern tools for administering and analyzing self-report measures. Conclusions forward a set of conditions for the use of self-report measures, which center on the role of theoretically-driven choices in both the selection of self-report measures and analysis of the data these measures generate.

Keywords: *self-report, self-regulated learning, motivation, strategies, strategy use*

Corresponding author: Peggy N. Van Meter, 226 Cedar Building, University Park, PA 16802, USA. Email: pnv1@psu.edu *DOI:* 10.14786/flr.v8i3.631.

Student learning can be understood through a lens of self-regulation, which explains learning as involving dynamic, cyclic processes that are both self- and goal-directed (Zimmerman, 1990). The self-directed component of this definition is central to understanding models of self-regulated learning (SRL) because these models posit that learning is influenced by the learner's own choices and abilities to apply effortful, effective learning processes. As such, the study of motivation and strategy use is a centrepiece in the study of SRL. If one is to understand a learner's choices and abilities, then one must understand the motives and strategic operations on which these rest. One must, that is, be able to answer questions such as, "What factors influence learners' motivational states?", "Which strategies do learners apply?", and "How do motivation and strategy use influence learning?". Although each article in this special issue contributes empirical evidence that advances our understanding of motivational and strategic processes and just how we might answer these questions, individually they differ with regard to the specific constructs of interest. Vriesema and McCaslin (2020), for example, report on the processes of identity formation in small group learning. Moeller et al.(2020) study the emotions and beliefs of interest for class activities while Rogiers, et al. (2020) examine how profiles of learning are associated with dynamic strategy use during text learning. Altogether then, these articles do give insights into a variety of SRL constructs associated with motivation and strategy use and each can be interpreted in the context of the corresponding construct-specific research literature. The focus of this special issue, however, is not on these constructs per se, but rather how these constructs are measured and how they are understood through the lens of that measurement. Specifically, the purpose of this special issue is to examine the use of self-report measures and address the challenge of determining "when and if self-report measures can contribute to our collective understanding of theory surrounding constructs." (Fryer & Dinsmore, 2020). The articles in this issue represent different ways of answering that call. Articles by Fryer and Nakao (2020), Iaconelli and Wolters (2020), and Chauliac et al. (2020) for example, adopt a measurement approach and focus on factors that can influence the reliability and validity of self-report data. Other articles; namely those by Durik and Jenkins (2020) and Moeller et al. (2020); explore methods to enhance the evidentiary value of self-report data. A final grouping of articles sought to establish the need for self-report data by demonstrating the benefits of using these instruments in pursuit of theoretically compelling components of learning. Included in this grouping are articles by Vriesema and McCaslin (2020), van Halem et al. (2020), and Rogiers et al. (2020).

Despite these differences, what unites these articles is shared attention to the set of central questions that drive this special issue. Specifically, author teams were tasked with addressing some combination of three questions that concern the utility of self-reports for the study of motivation and strategy use. These central questions ask about (1) the alignment of self-report methodology and theoretical conceptualizations of constructs, (2) the influence of self-report methodology on the interpretation of study results, and (3) the connection between self-report methodology and analytic choices. The articles in this issue present data obtained through particular programs of research and, as such, each article offers some particular view on the answers to these questions. The goal of this commentary is to look across those specifics and offer a more synthetic perspective; one that draws across constructs and methodologies to highlight themes around these questions and draw conclusions about what this body of articles suggests for the future of self-report use. Toward that end, my comments will admittedly overlook differences with regard to the specific constructs represented in this set of papers and instead treat each as representative of the set of constructs associated with SRL. The remainder of this commentary is organized around the three central questions and addresses some of the major themes that emerged from the articles in this special issue.

In what ways do self-report instruments reflect the conceptualization of the constructs suggested in theory related to motivation and strategy use?

On the face of it, this is a rather straightforward question about an aspect of construct validity. That is, do the measures align with, and therefore reflect, the theoretical conceptualizations of the constructs (Edwards & Bagozzi, 2000)? Construct validity is critical to the relationship between measurement and theory because it is measurement that provides the operational definition of a

construct. Whereas theoretical descriptions of a construct may be abstract and difficult to pin down, an operational definition is the specific way in which the construct is measured, including the exact prompts to which participants respond and the ways that data is collected. Consequently, if one wants to know what is meant by theoretical terms such as identity formation (Vriesema & McCaslin, 2020) or interest (Fryer & Nakao, 2020), one need only look to how those constructs are operationalized. In this regard, establishing this aspect of construct validity requires three elements (1) a clear articulation of the theoretical conceptualization, (2) a clear articulation around the measurement methodology, and (3) a coherent mapping between the conceptualization and the methodology. Efforts toward establishing construct validity can also feed into a cycle of theory and measurement refinement. That is, confidence in the validity of a measure is increased when obtained data behave in theoretically consistent and predictable ways, but innovations in measurement can also reveal evidence of phenomena that stimulate refinement and development of theoretical accounts.

The articles in this special issue provide a number of examples of how this form of validity can be established when using self-report measures. Most specifically, the authors achieve this by carefully and explicitly defining the constructs of interest in the context of guiding theoretical frameworks, and tying these definitions to the measurement instrument. While several articles provide examples of how this can be done, just two will be presented as illustrations here. First is the study by Rogiers et al. (2020), in which they examined the text learning strategies of middle school students. The purpose of this study was to "fully map and understand individual students' learning" (p. 1) using the research context of students studying to learn from an expository text to engage in this mapping. SRL is the theoretical framework that guides this research and, consistent with the definitions used throughout this issue, Rogiers et al. defined SRL as involving adaptive, flexible strategy use in dynamic, iterative phases. Further, Rogiers et al. stated that there are individual differences in how learners employ strategies and in their perceptions of this strategy use. Most central to the theoretical conceptualization, Rogiers et al. also reasoned that these individual differences could provide insight into the dynamic, adaptive ways that learners employ strategies during learning. Their use of two different self-report measures follows from this conceptualization. First, participants thought aloud while engaged in the text learning task with the resulting protocols revealing of the dynamic strategic processes employed during the task. Second, after reading, participants completed a self-report survey, which queried the task-specific cognitive and metacognitive strategies used during study. This survey measure identified meaningful individual differences and served to group participants into different profiles of strategy use (e.g., integrated strategy user). The value of both forms of self-report data was realized by using the profiles of strategy use to guide interpretation of think aloud data. In brief, consistent with theoretical conceptualizations, Rogiers et al. were able to use self-report measures to demonstrate that different types of strategy users employ dynamic SRL processes in different ways.

A second example of how articles demonstrate the connection between conceptualizations of a construct and self-report measures of that construct can be found in Moeller et al.'s (2020) study of situational interest in a college course. This article defined interest as a motivational and emotional state that fluctuates over time, and measured these fluctuations as situational expectancy and task value (i.e., expectancy-value; Eccles, & Wigfield, 2002). Moreover, the authors argued that, at any point in time, these states are a function of (1) stable personal traits, (2) situational personal perception, and (3) objective components of the situation. In order to follow from this theoretical conception then, measures of interest must capture and distinguish all three sources of this variance. The use of a self-report interest survey, which was administered periodically during class, was a logical choice in this context because survey responses allow the capture of individuals' perceptions. It was the manner in which Moeller et al. employed the survey, however, that permitted the strong connection between the theoretical conceptualization of situational interest and the self-report measure. While the reader is referred to that article for a full explanation of the methodology, a short summary here will suffice: Course students completed the survey at multiple time points with multiple students intentionally sampled at each time point. This sampling pattern then permitted examination of both objective evaluations (i.e., group means) and personal perceptions (i.e., deviation from the mean). Ultimately, the use of the self-report measure was validated when Moeller et al. were able to parse the variance in individuals' time-point interest reports into the three theoretically predicted sources of variance.

In sum, the articles in this special issue demonstrate that self-report measures can not only reflect conceptualizations of constructs, they can do so in theoretically compelling ways. These efforts toward construct validity feed the mutually reinforcing cycle of theory development and methodological refinements. Rogiers et al.'s finding of relations between profiles based on learners' perceptions of strategy use and their dynamic application of those strategies, that is, furthers understanding of individual differences and SRL. At the same time, Moeller et al.'s study advances theory regarding the personal and objective sources of situational interest; a motivational construct central to understanding SRL. In the context of this special issue, however, where the challenge is to determine when and if self-report data contributes to the understanding of constructs, there is another layer to the question of how self-report reflects theoretical conceptualizations. Specifically, in this context, it is not sufficient to show that some measurement choice is consistent with theoretical definitions or even that the self-report data accounts for some theoretically interesting variance. Instead, this task calls on us to consider when and if self-report data provides insight into some phenomenon that is not gained by another measurement approach. In other words, we are challenged to show not only that self-report measures can reflect conceptualizations of motivation and strategy use, but also that some self-report methodology is uniquely suited to doing so.

A partial response to this challenge can be obtained by pointing back to the constructs of interest. Specifically, when the construct of interest is a learner's perception of intra-psychic states (e.g., beliefs, motivations), then it is sensible to suggest that the best way to uncover these perceptions is to ask the learner (Fryer & Nakao, 2020). In addition to this argument, however, articles in this special issue lay out an even more convincing reason for using self-report measures. Namely, self-reports are a justifiable measurement tool because data from these measures offer unique explanatory power when it comes to understanding motivation and strategy use. Again, two studies from this special issue can be used to illustrate this point. The first example is the study by Vriesema and McCaslin (2020), which sought to understand the processes of identity formation in small group settings. Guided by a co-regulation theoretical model, the authors collected self-report data on students' perceptions of how they engaged with the members of their group as well as pre- and post- anxiety and emotional adaption profiles. Observational data of group interactions was also collected and analyzed to show the actual interaction pattern that took place in the groups over a series of six lessons. The analysis of this data demonstrates that more is learned about identity formation and co-regulation from both self-report and observational data than from either source alone. While pre-group self-reports of emotional adaptation were predictive of some co-regulation styles, for example, certain co-regulation styles were predictive of post-group emotional adaptation profiles.

The value of self-report methodology is also demonstrated in the study by van Halem et al. (2020), which shows that data obtained from these measures offers unique explanatory insights. In this study, trace data was collected over a period of eight weeks as students in a college statistics course accessed online course resources. Using the theoretical framework of SRL, the authors point out that these trace data provide insights into behavioral aspects of learning; these traces are "observable evidence of particular cognitions…where a cognitive process is applied" (p. 3) At the same time, however, these traces do not indicate just what those cognitive processes are. One student, for example, may access some resource because it covers content from a class that was missed while another student may access that same resource because they did not understand the content when it was covered in class. To gain insight into the processes underlying these behavioral traces, van Halem et al. had participants complete a self-report survey of SRL behaviors (i.e., Motivated Strategies for Learning Questionnaire; MSLQ; Pintrich et al. 1993) in the fourth week of the course. At the end of the course, analyses showed that, when both trace and self-report data were included, some MSLQ sub-scales accounted for variance in grades above and beyond that accounted for by the behavioral data. In short, like the research on identity formation in small groups, this study shows that a self-report measure can explain important aspects of motivation and strategy use that would not be captured in the absence of the measure.

In summary, the research teams represented in this special issue demonstrate three specific ways in which self-report instruments reflect theoretical conceptualizations of motivation and strategy use. First, across the set of articles, one can see that self-report instruments and methodologies can operationalize constructs in theoretically consistent ways. Second, these measures can generate data that

not only behaves in theoretically predictable ways, but also offers refinements to the conceptualization of constructs. Third, self-report measures can reflect conceptualizations by capturing patterns and variance in motivation and strategy use that are not obtained through other means. While these answers justify the use of self-report from a conceptual standpoint, they cannot be completely disentangled from more specific methodological choices associated with the use of self-report. The two remaining questions posed by this special issue provide the opportunity to address some of these points.

How do the interpretations of self-report data influence interpretations of study findings?

This second question, which asks how the interpretations of self-report data influence the interpretations of study findings, is similar to the first question in that it can be understood as addressing an aspect of validity. Namely, validity is not determined by some measure itself but rather by the degree to which the interpretations drawn from the scores on that measure can be justified (Messick, 1995). In this respect, the interpretations of a study's findings are valid when the data sources on which those findings are based have been interpreted in valid ways. This logical argument then calls for a particular view on the question that frames this section: To understand how self-report data influences the interpretation of study findings, we must understand the factors that influence the reliability and validity of the data derived from these measures. Also similar to the previous section, there are two different perspectives we can take on this question. The first perspective concerns the factors that may influence the reliability and consequently, the validity of scores from self-report measures. This perspective is primarily concerned with potential sources of error in self-report measurement of motivation and strategy use. The second perspective takes a more conceptual view and considers the ways in which the methodologies of collecting self-report data influence the interpretations of that data. This perspective draws attention to the broader theoretical and contextual factors that influence how scores can be interpreted.

With respect to the first perspective, several studies in this special issue examine sources of error in self-reports and how those sources can be understood or reduced. One potential source of error, which is examined in the study by Fryer and Nakao (2020), is the format of the response scales and interfaces used to record participant's responses. This examination was prompted by the body of work on survey instruments, which suggest that the response scales themselves can impact the nature and reliability of scores. Participants in this study were graduate students enrolled in a course on teaching and, throughout the course, these participants responded to surveys assessing their interest in class activities. To examine response formats as a potential source of measurement error, this study had participants complete self-report surveys that asked the same questions, but used four different interfaces: labelled categorical scales, visual analog scales (VAS), swipe, and slider. These surveys were administered at six time points throughout the semester so that all participants responded using each of the interfaces and, at any one time point, all four interfaces were used. This design permitted comparisons across the different interfaces to determine if any significant differences in response patterns could be tied to differences in the interface. On the whole, the results suggest that response interfaces are not a significant source of error. Each of the measurement methods yielded acceptable levels of reliability and there were no differences in either the mean scores across the measures within the six time points or differences in the factor structures of the measures. Although details in the findings lead Fryer and Nakao (2020) to suggest that the Swipe method shows promise and the VAS method is the weakest, the totality of the data indicates that scores obtained from each of the response formats and interfaces can be validly interpreted.

Another potential source of error, one that has been suggested throughout the history of self-report surveys, is insufficient effort on the part of respondents. According to this view, the results of self-report surveys are tainted by participants who either do not put forth the cognitive effort to answer survey questions or bias the results by responding in unserious ways. Two articles in this special issue address this concern by examining data related to participants' survey response patterns. First, as part of a larger study, Chauliac et al. (2020) collected eye tracking data while college students responded to a task-specific survey on processing strategies (i.e., Inventory of Learning Styles; Vermunt & Donche, 2017). The time and frequency of eye fixations on any given question were interpreted as indicators of effort while the consistency of an individual's within-scale responses were taken as an indicator of

within-person reliability. Analyses showed a relationship between fixations and response variability wherein participants who spent the most time on an item were also most likely to show only small degrees of variation in item responses. In other words, these participants showed patterns indicative of reliable responding. By contrast, participants who spent the least amount of time were most likely to either select the same categorical response for each scale item or show extreme variability; i.e., poor reliability.

Chauliac et al.'s (2020) finding are complimented by the research of Iaconelli and Wolters (2020), which also examined indicators of insufficient effort responding, but extends this work by presenting techniques to detect these participants in large scale data collections. This study involved nearly 300 college students in a course designed to improve their SRL and, at three different times in the course, all participants completed self-report surveys tapping into their dispositions, beliefs, and behaviors. Consistent with Chauliac et al. (2020), Iaconelli and Wolters posit that the validity of a measure is threatened if respondents exert too little effort while answering questions. Further, they posit that these response patterns can be detected by examining indicators of effort (i.e., time) and consistency in response patterns. While the reader is referred to the article itself for details on these indicators, there are three main conclusions relevant here. First, there are some participants who show insufficient effort. But, two, these participants comprise only a small percentage of the total sample and their inclusion, at least in a large data set, does not significantly alter either the mean or the internal consistency of the data set. Third, although each of the three surveys had some participants who gave insufficient effort, this did not appear to be a stable individual difference. Instead, if a participant did exhibit insufficient effort, this tended to occur on only one of the three surveys.

As summarized here, the articles in this issue report evidence that self-report surveys can and do provide reliable indicators of variables associated with motivation and strategy use. Neither the response format nor a lack of respondent effort introduced sufficient error variance to question the interpretations that are drawn from these instruments. Under these conditions then, we can conclude that study findings based on these self-report measures can be interpreted in the intended ways. The second perspective on this question of how self-reports influence study findings, however, encourages us to look at a broader set of factors that influence the validity of self-report data interpretations. These broader factors include the totality of the context in which the measure is administered including theoretically-motivated methodological factors. To illustrate this, consider the article by Durik and Jenkins (2020), which ultimately concludes that the person-domain context must be taken into account when interpreting the results from self-report surveys of learner interest. In a pilot study and two experiments, college students responded to self-report surveys that assessed their interest in different domains (e.g., math and psychology) and also indicated the likelihood that they would pursue future learning in these domains (Study 1 and 2). The purpose of this research was to explore how self-report can be used to better explain the relationship between interest and behavior and, toward that goal, Durik and Jenkins had participants also respond to questions gauging their certainty in provided interest ratings. Factor analyses showed that interest and certainty comprise separate factors, indicating that respondents are able to distinguish these two beliefs. Analyses also showed that the level of certainty moderated the relationship between reported interest and future behavior with the interest-behavior relationship markedly stronger for participants with high levels of certainty. Altogether, the data presented in this article shows that, at least in the study of interest, (1) participants' ability to provide accurate, predictive self-reports depends on how certain they are about these reports and (2) certainty varies with exposure to the domain. Considering this in light of the question of how self-report influences the interpretation of study findings, this research highlights the need to attend to the broader context in which the measure is administered; in this case, the context of the person-domain relationship.

Another methodological factor that emerged as important to the interpretations of study results is the timing of self-report administration. Although there are different types of self-reports possible, each requires participants to respond to some query on the basis of their memory for the relevant information (See the articles by Chauliac et al. and Iaconelli & Wolters, this issue for a discussion of these models) and each is administered prospectively, concurrently, or retrospectively. In this respect, self-report responses provide a snapshot in time (Durik and Jenkins, 2020). Yet, because effective learning processes are understood as dynamic, flexible, and adaptive; a challenge to the use of self-

report data is the need to show how a snapshot can shed light on active motivational and strategic operations. This special issue presents one possible answer to this: Researchers can enhance the validity of interpretations by attending to the timing in which self-report measures are administered and incorporating this timing into the interpretation of study findings.

To illustrate this point, consider the study by van Halem et al. (2020) in which trace data was collected from students as they accessed online materials throughout an eight-week statistics course. Participants also completed a self-report measure of SRL in the fourth week (i.e., MSLQ). As described previously, study results showed that both the trace and self-report data accounted for variance in students' final course grades. Additionally, however, van Halem et al., also examined relations between trace and self-report data for each of the eight course weeks and found that the relationships were the strongest in the weeks preceding completion of the self-report survey and weak in the periods thereafter. In short, the timing of the self-report measure influenced the nature of the resulting data and thus, must be incorporated into the interpretation of study findings: Self-report can be an accurate snapshot of students' memories for what they have done in a course but are not necessarily prognosticators of future behavior, at least not with respect to SRL as measured by the MSLQ.

The influence of timing in the administration of self-report measures is also demonstrated in the study by Rogiers et al. (2020). As explained in the previous section, middle school students in this study thought aloud while reading expository text and, immediately after reading, completed a self-report survey of the strategies used. In this respect, both concurrent and retrospective self-report measures are used with the retrospective survey placed close in time and in direct reference to the just-completed SRL event. Again, this timing influences how the data can be interpreted. First, because the survey immediately followed the SRL event, results can be interpreted as valid representations of learners' perceptions of their strategy use and; second, concurrent think alouds reveal the pattern in which strategies were used. Finally, Rogiers et al. were able to use the profiles that emerged from retrospective self-reports to guide data mining and uncover differences in how individuals deploy SRL processes. The timing matters here because it is the time-based relationship of the two self-report measures that permit the data and study findings to be interpreted in this way.

The two methodological factors covered here, person-domain relations and timing, are just two of the contextual factors addressed in this special issue that should be considered when interpreting study findings. Vriesema and McCaslin's research on identity formation in groups, for example, demonstrates that this development must be understood in the context of the specific group's dynamics; i.e., individuals are nested in groups. Exactly how data is collected should also be taken into consideration. Iaconelli and Wolters (2020) show this in their examination of insufficient effort responding. Recall these authors found that, while insufficient effort responding does occur, these occurrences have a negligible effect on the data set. These authors, however, were careful to point out that the surveys were completed as part of homework assignments in participants' course on SRL. As a result, it is possible that insufficient effort responding was infrequent in this study because participants had a high degree of investment. Higher, that is, than one might expect from study participants who complete a survey only to receive course extra credit for study participation (e.g., Durik & Jenkins, 2020).

Taken as a whole, the current set of articles provide at least two important insights into the ways that the interpretation of self-report data can and should be used to interpret study findings. First, self-report data sources can be trusted to provide reliable and valid indicators of studied constructs. Although there is some error in these measures, evidence culled from these studies provide confidence that this is no greater a problem for self-report measures than other types of data collection methods that rely on human responses. Second, the broader contextual and methodological factors of measurement administration matter. Although the studies reported here shed light on some of these factors, no doubt there are many more that warrant attention. As a summary though, one can conclude that the methodology around the administration of self-report measures influences the interpretation of resulting data and consequently, influences the interpretation of study findings.

How does the use of self-report constrain the analytical choices made with that self-report data?

This final question can be understood to specifically address data analytic concerns rather than issues related to construct conceptualizations and the interpretation of findings covered in the first two questions. With these boundaries in mind, the short answer to this question is that the self-report nature of this data does not place constraints on analytic choices above and beyond what must be considered with other data sources; constraints such as scales of measurement, distributions, and floor or ceiling effects. Indeed, what is most striking in relation to this question are the creative and innovative ways in which self-report data can be collected and analyzed. Of course, it has long been argued that a chief benefit of self-report data is the ability to collect data from large sample sizes and this benefit is only increasing with technological advances in digital delivery systems (Fryer & Nakao, 2020). Beyond this ease, however, the articles in this issue highlight two valuable connections between the use of self-report data and subsequent analytic choices.

The first connection that emerged is how advances in both measurement delivery and statistical analytic tools permit self-report data to be collected and analyzed in increasingly creative and sophisticated ways. As authors in these special issue articles have noted, self-report measures have traditionally been delivered in paper-and-pencil form and resulting data treated in aggregated, variable-centered ways: Group means attest to some agreed upon (i.e., averaged) descriptor of a construct and scores are treated to traditional forms of comparisons and correlations. By contrast, today's researcher has access to much more sophisticated tools to deliver measures as well as to parse variance and model data patterns. Consider, for example, the study by Moeller et al. (2020) that examined both personal and objective contributions to fluctuating states of interest. Thus far, this commentary has described this research in terms of the studied construct and empirical findings. An examination of the study methods, however, illustrates how innovation in the delivery and analysis of self-reports is expanding our understanding of motivation and strategy use in SRL. Specifically, these authors leveraged online delivery mechanisms and innovative experience sampling methods to collect the self-report data from groups of participants in context and over time. Once collected, the application of cross-classified multilevel modelling permitted participants' time-point self-report data to be parsed into the three sources of variance that were predicted by the theoretical framework of situational interest used in this study. In short, Moeller et al. were able to apply modern research tools to the collection and analysis of data in a manner that advances theoretical understanding. Moeller et al.'s work is not the only illustration of the ways that self-report data can be meaningfully analyzed given the tools currently available. The studies by both Iaconelli and Wolters (2020) and Fryer and Nakao (2020), for example, show how the online delivery of self-report measures can yield not only participants' responses but also indicators of invested effort (i.e., time). Others took advantage of recent techniques to detect patterns within data sets and used these methods to mine for dynamic, iterative, SRL cycles (Rogiers et al., 2020); classify participants according to profiles of individual differences and group co-regulation dynamics (Vriesema & McCaslin, 2020); and disentangle the relationships between interest, certainty, and behavior (Durik and Jenkins, 2020).

Altogether, this special issue shows that the choice to use self-report data opens the door to a great number of analytic choices, but the most promising of these may be the use of self-report data alongside other, complimentary data sources. As previously summarized, self-report data can have unique explanatory power when combined with other data sources in the study of motivation and strategy use. That previous discussion, however, was narrowly focused on how self-report reflects conceptualizations of theoretical constructs, and did not address methodological and analytic dimensions of this point. With respect to the current question though, one can see significant potential in the use of self-report measures in conjunction with additional measures of motivation and strategy use. For instance, self-report measures can play an important role in mixed methods SRL research in which qualitative process data can be combined with quantitative scores derived from self-report surveys. Just such an approach is demonstrated in the studies by both Rogiers et al. (2020) and Vriesema and McCaslin (2020) in which qualitative process data was collected and coded in addition to the administration of self-report surveys. Ultimately, these data sources were combined to shed light on how self-reported individual differences related to SRL processes in either individual (Rogiers et al., 2020) or group (Vriesema & McCaslin, 2020) settings. In addition to these two studies, articles in this special issue show other ways of combining self-report survey responses with additional forms of process data

such as eye fixations (Chauliac et al., 2020), trace data (van Halem et al., 2020), and response times (Iaconelli & Wolters, 2020).

In sum, this section can be closed by returning to the answer offered at the opening; namely, the self-report nature of some data does not place constraints on analytic choices above and beyond what must be considered with other data sources. Instead, what does constrain both the choice of measures and data analysis methods, are the theoretically-based conceptualizations of the construct and the questions that drive the research. As the articles in this issue show, self-report data can be analyzed in a wide variety of ways with innovations paving the way for breakthroughs in both measurement and theory. Certainly, one must be concerned with the psychometric properties of scores and the match between the data set and the assumptions of a particular analysis. Beyond these constraints, however, self-report data has, and can continue to be, analyzed in ways that yield relevant insights into the individual differences and processes of motivation and strategy use.

Conclusion and Final Remarks

The articles in this special issue shed light on a number of theoretical constructs associated with motivation and strategy use, but the main objective of this collection is to examine the self-report methodology used to study these constructs. The task for the articles in this issue, including this commentary, was to use three organizing questions to "determine when and if" (Fryer & Dinsmore, 2020) self-report measures positively contribute to the study of theoretical SRL constructs. This final conclusion section focuses on this task by considering first, the question of "if" self-report measures can contribute followed by the question of when this might be true.

The question of "if" self-report measures can be used calls for answers to two relatively straightforward questions: (1) Is there evidence that scores on self-report measures can be reliable and valid indicators of motivational and strategic constructs? and (2) Is there evidence that self-report measures provide explanatory power in the study of motivational and strategic constructs? Across all of the articles in this special issue, the answer to both of these questions is, "Yes". One bit of evidence in support of this affirmative response is found in demonstrations that self-report measures yield scores with acceptable reliability and adequate psychometric properties. Iaconelli and Wolters (2020), for example, showed that insufficient effort responding had little impact on a full data set and Moeller et al. (2020) showed how theoretically-driven analysis can increase the amount of variance explained in self-report data. Additional evidence from this set of articles comes from the repeated demonstrations that self-report measures play an important role in capturing and understanding theoretical constructs. In short, this body of research advances our understanding of motivation and strategy use and this is due, in large part, to the use of self-report measures. From the data presented in these studies, we learned about SRL phenomena such as situational fluctuations in motivational states, the relationship between group dynamics and identity formation, individual differences in the dynamic application of strategy use, and the role of domain exposure and certainty in understanding the influence of interest on behavior. In short, the research in this special issue supports the conclusion that self-report measures do indeed have an important role to play in the study of motivation and strategy use. And, this is true whether one is focused specifically on psychometric measurement properties or theoretically-driven conceptualizations.

The second part of our task, the task of determining "when" self-report measures contribute to the study of motivation and strategy use, raises questions about the conditions under which self-report may or may not be appropriate. Indeed, the articles in this special issue raised concerns about several of the limitations of self-report measures. For example, because self-report measures capture a snapshot of a learner's perceptions, these instruments may be better at explaining a learner's past than predicting that learner's future (e.g., van Halem et al., 2020). Likewise, when self-reports are in the form of surveys, they capture variance associated with motivation and strategy use, but do not effectively capture dynamic aspects of SRL (e.g., Vriesema & McCaslin, 2020). Finally, like any other measure, self-reports are not immune to potential sources of error such as individual differences (Moeller et al., 2020; Durik and Jenkins, 2020) and insufficient effort responding (e.g., Chauliac et al., 2020). These limitations notwithstanding, it is possible to draw conclusions about when self-reports are likely to advance the

study of SRL constructs. Specifically, there are three conditions under which self-report measures can be effectively used: (1) when the measure aligns with theoretically-driven conceptualizations of the construct, (2) when measure selection is driven by alignment with theoretically-driven research questions, and (3) when measure administration, data analysis, and results interpretations are grounded in theoretically-driven choices. In sum, self-report measures can contribute to the study of motivation and strategy use when a close coupling of the measure and relevant theory allows for a mutually beneficial cycle of refinement and development.

With these recommendations in mind, I will close with one final thought that points out the primary weakness of this commentary; namely, the choice to synthesize across the specific constructs studied in each article and group them under the broad umbrella of SRL. While this choice permitted general conclusions to be drawn about the use of self-report in the study of motivation and strategy use, it also meant that attention was not paid to possible construct-measurement interactions. That is to say, interactions in which the conditions for how and when self-report measures are best used vary according to the construct under study. Reading this set of papers, for example, raises a number of interesting questions about these possible interactions; e.g., whether the degree of certainty influences the prognostic abilities of self-reported strategy use in the same way that it influences measures of interest, if the rates of insufficient effort responding are consistent across SRL constructs, how the methods used to evaluate the effects of classroom activities on interest could be used to evaluate how those activities stimulate strategic processes. Despite the lack of attention given here to possible construct-measurement interactions such as these, their exploration offers a direction for future research. Carrying out this work has the potential to shed light on not only the use of self-report measurement tools, but also the theoretical conceptualizations in which they are grounded.

Keypoints

- Self-report measures can accurately and constructively reflect theoretical conceptualizations of SRL constructs.
- Data derived from self-report measures can provide reliable and valid indicators of motivation and strategy use.
- The application of modern research tools to the use of self-reports can lead to breakthroughs in both SRL measurement and theory.
- Self-report is most effectively used when it is closely aligned with theory.

References

Chauliac, M., Catrysse, L., Gijbels, D., & Donche, V. (2020). It is all in the surv-eye: Can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research, 8*(3), 26–39. https://doi.org/10.14786/flr.v8i3.489

Durik, A., & Jenkins J. (2020). Variability in certainty of self-reported interest: Implications for theory and research. *Frontline Learning Research, 8*(3), 86–104. https://doi.org/10.14786/flr.v8i3.49

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109-132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155. https://doi.org/10.1037/1082-989X.5.2.155

Fryer, L. K., & Dinsmore, D. L. (2020). The promise and pitfalls of self-report: Development, research design and analysis issues, and multiple methods. *Frontline Learning Research*, *8*(3), 1–9. https://doi.org/10.14786/flr.v8i3.623

Fryer, L. K., & Nakao, K. (2020). The future of survey self-report: An experiment contrasting Likert, VAS, slide, and swipe touch interfaces. *Frontline Learning Research*, *8*(3), 10–25. https://doi.org/10.14786/flr.v8i3.501

Iaconelli, R., & Wolters, C. A. (2020). Insufficient effort responding in surveys assessing self-regulated learning: Nuisance or fatal flaw? *Frontline Learning Research, 8*(3), 105–127. https://doi.org/10.14786/flr.v8i3.521

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741. https://doi.org/10.1037/0003-066X.50.9.741

Moeller, J., Dietrich, J., Viljaranta, J., & Kracke, B. (2020). Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Frontline Learning Research, 8*(3), 63–85. https://doi.org/10.14786/flr.v8i3.529

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*(3), 801-813. https://doi.org/10.1177/0013164493053003024

Rogiers, A.; Merchie, E., & van Keer, H. (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research, 8*(3), 40–62. https://doi.org/10.14786/flr.v8i3.527

van Halem, N., van Klaveren, C. P. B. J., Drachsler, H., Schmitz, M., & Cornelisz, I. (2020). Tracking patterns in self-regulated learning using students' self-reports and online trace data. *Frontline Learning Research, 8*(3), 142-164. https://doi.org/10.14786/flr.v8i3.497

Vermunt, J. D., & Donche, V. (2017). A learning patterns perspective on student learning in higher education: state of the art and moving forward. *Educational Psychology Review*, *29*(2), 269-299. https://doi.org/10.1007/s10648-017-9414-6

Vriesema, C. C., & McCaslin, M. (2020) Experience and meaning in small-group contexts: Fusing observational and self-report data to capture self and other dynamics. *Frontline Learning Research, 8*(3), 128–141. https://doi.org/10.14786/flr.v8i3.493

Zimmerman, B. J. (1990) Self-Regulated Learning and Academic Achievement: An Overview, *Educational Psychologist, 25*(1), 3-17, https://doi.org/10.1207/s15326985ep2501_2

Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol. 8 No. 3 special issue (2020) 186 - 194
ISSN 2295-3159

Commentary: Self-Report is Indispensable to Assess Students' Learning

Reinhard Pekrun[a]

[a]University of Essex, United Kingdom & Australian Catholic University, Sydney, Australia

Abstract

Self-report is required to assess mental states in nuanced ways. By implication, self-report is indispensable to capture the psychological processes driving human learning, such as learners' emotions, motivation, strategy use, and metacognition. As shown in the contributions to this special issue, self-report related to learning shows convergent and predictive validity, and there are ways to further strengthen its power. However, self-report is limited to assess conscious contents, lacks temporal resolution, and is subject to response sets and memory biases. As such, it needs to be complemented by alternative measures. Future research on self-report should consider not only closed-response quantitative measures but also alternative self-report methodologies, make use of within-person analysis, and investigate the impact of respondents' emotions on processes and outcomes of self-report assessments.

Keywords: self-report; emotion; motivation; metacognition; self-regulated learning

Reinhard Pekrun, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, Email: pekrun@lmu.de
*DOI:*10.14786/flr.v8i3.627

Introduction

Self-report is indispensable for any more nuanced assessment of mental states. While it is possible to examine general physiological properties of thought and affect using brain-imaging, and their consequences through performance tests and behavioral observation, assessing the contents and complex cognitive processes involved in human thinking, emotion, and motivation requires self-report. As such, self-report was a primary assessment method in psychology and education from early on, and it continued to be a primary method throughout all developmental phases in the history of these disciplines, even in the prime time of behaviorism early in the 20th century. However, self-report also has limitations. Self-report is restricted to processes that are accessible to consciousness; is typically limited to assess contents that can be verbally described; can be subject to various biases; and is always lagging behind the processes it aims to assess, even if only for seconds, which implies that it lacks the temporal resolution needed to capture the real-time dynamics of learning.

Given these problems, it is important to critically scrutinize the power of self-report methods to capture the constructs they intend to assess, and to develop strategies to improve their validity. The papers in this special issue are excellent examples for both directions. Specifically, all eight papers examine the validity of specific self-report instruments relative to proposed distributions of scores and relations with other variables. In addition, two of the papers also explore ways to improve validity. In the following sections, I first address the nature of self-report and its advantages and drawbacks. Next, I discuss the advances in analyzing and improving the validity of self-report measures that are represented in the contributions to this special issue. In conclusion, I outline three directions for future research.

1. What is self-report?

Self-report uses participants' verbal responses to assess their cognition, emotion, motivation, behavior, or physical state. When thinking about self-report, what often comes to mind first is structured questionnaires measuring some kind of personality trait. However, while structured questionnaires are used frequently, the most commonly employed self-report instrument likely is the clinical interview, which typically has a very different format as compared with closed-response questionnaires. By implication, to judge self-report, it is important to consider that this method can take very different forms. Self-report can be structured or unstructured; retrospective or concurrent; oral or written; qualitative of quantitative; one-dimensional or multi-dimensional; paper-and-pencil or online; and can comprise single or multiple items (see Pekrun & Bühner, 2014, for an overview). As such, self-report not only includes structured multi-item questionnaire scales, but also open-ended interviews, single-item momentary reports, unstructured think-aloud protocols, etc. While all these methods share properties of relying on participants' ability to self-assess and report about the variables under investigation, they differ vastly in terms of structure, temporal resolution, and metric used. As such, it is important to keep in mind that any findings on the validity of self-report instruments, and on ways to improve it, may be specific to some variant of self-report and not be generalizable to other variants.

In the current special issue, all of the eight contributions consider multi-item questionnaire scales using closed formats (i.e., defined items and response options). Rogier et al. (2020) additionally included a think-aloud protocol. As such, with this exception, the contributions focus on quantitative, structured self-report measures. Such measures are well suited to answer quantitative research questions that are defined a priori. However, they are less suited to answer exploratory questions and to gain a more nuanced picture of respondents' subjective world of multi-layered thoughts and perceptions, which can transcend researchers' prior conceptions as represented in closed-format scales. For such purposes, qualitative self-report methods are needed. Overall, to make progress in research on learning and instruction, it is often useful to employ a mix of quantitative and qualitative

self-report, with qualitative methods used to explore new territory and gauge in-depth explanations, and quantitative method to test theoretical hypotheses in more rigorous ways (see, e.g., Pekrun et al., 2002).

2. Benefits and drawbacks of self-report

Self-report has clear advantages. First, in contrast to other types of assessment, self-report allows assessment of all types of psychological processes. Observation can assess visible behavior, achievement tests can measure cognitive performance, neuro-imaging the activation of brain areas, and physiological analysis the arousal of peripheral systems, but self-report can be used to assess all of the affective, cognitive, physiological, and behavioral processes that are part of self-regulated learning – all of these processes can be represented in the human mind and can be reported accordingly. Second, self-report can render a more differentiated assessment of human thinking than any other method. As such, for a nuanced description of emotions, motivation, and metacognition during learning, self-report is needed. Third, self-report is more economical than other methods. Self-report may be the only method applicable in some types of studies, such as large-scale student assessments.

Self-report also has disadvantages. As noted, self-report is limited to the assessment of processes that are accessible to consciousness. Responses that cannot be represented mentally need to be assessed with other methods. Another important limitation is the use of language (although self-report can also employ non-verbal communication). Research has shown that terms describing psychological processes tend to be used in consistent ways across languages (e.g., Fontaine et al., 2013), but there can nevertheless be differences in semantic understanding across cultures and learners. By implication, measurement equivalence of self-report instruments across groups should not merely be assumed but needs to be established empirically. Furthermore, limitations result from the fact that self-report is under respondents' control. Whereas it may be difficult to alter one's level of physiological activation, reports about perceived activation can easily be changed. As such, depending on motivation and preferences for response options, self-report can be subject to various response biases, such as social desirability.

Finally, self-report is also subject to memory biases. This is especially true for retrospective self-report that is administered at a later point in time and requires recollection of information from autobiographical memory, but is also true for state self-report asking respondents how they feel or what they think right now – self-report is lagging behind the phenomena it captures, even if only for seconds. As such, self-report inevitably lacks the temporal resolution needed to examine the real-time dynamics of psychological processes. This is even true for momentary methods such as experience sampling or think-aloud protocols as used in the contributions by Dietrich et al. (2020) and Rogier et al. (2020). Even these methods cannot reach the temporal granularity of concurrent physiological or behavioral-observational methodologies. As such, self-report needs to be complemented with other methods for many research purposes. For making progress in research on learning, multi-channel assessments of motivation, emotion, and metacognition including self-report along with observational and physiological methods as well as behavioral trace data are especially promising (Azevedo et al., 2018; Lajoie et al., in press).

3. Examining the validity of self-report measures

Six of the eight contributions to this special issue focus on examining the validity of (quantitative) self-report measures and developing methods to examine validity. Iaconelli and Wolters (2020) investigated the impact of insufficient effort in responding on university students' self-report scores for their beliefs and behaviors during self-regulated learning. Self-report was assessed as part of students' coursework. Rates of insufficient effort were low, and reported relations between variables were robust against including students with insufficient effort, suggesting that inattentive responding

does not represent a major threat to validity (at least under the situational conditions of the study).

Rogiers et al. (2020) examined secondary school students' retrospective self-report of learning strategies used while learning from a text in combination with think-aloud data obtained during the same session. The data from the retrospective self-report were used to classify different types of learners, and the findings show that these types differed systematically in their learning process as assessed through the think-aloud protocol, thus attesting to the convergent validity of these two – very different – types of self-report instruments.

Extending the perspective beyond individual learning, Vriesema and McCaslin (2020) used self-report to assess secondary school students' general test anxiety, their attitudes towards school, and their behavior and emotions during group work in mathematics. There were clear links between self-reported behavior and emotions related to the group work situation, but less so with the test anxiety measure. These findings are consistent with the specificity matching principle (see, e.g., Swann et al., 2007): Variables show stronger relations when being matched in terms of situational specificity than when not being matched, and the present results suggest that self-report measures can demonstrate validity when attending to this principle.

Van Halem et al. (2020) used the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1991) as well as online trace data to assess undergraduate students' motivation and learning strategies during a statistics course. There was no direct conceptual match between the MSLQ and online trace data constructs. Nevertheless, there were substantial relations between time investment as assessed by the MSLQ, on the one hand, and trace data, on the other, thus demonstrating convergent validity of the two types of measures. Furthermore, both the MSLQ scores and the online trace data contributed to explaining students' course performance, thus supporting predictive validity for both types of measures.

Dietrich et al. (2020) used experience sampling methodology (ESM) with a two-item measure of situational interest to capture the developmental dynamics of university students' interest in a series of lectures over one semester. In contrast to traditional ESM designs, a fixed rather than random schedule of assessments was used, which facilitated aggregation of assessments across participants. The findings of cross-classified multilevel analysis show that there was substantial variation of interest scores between students as well as within and between lectures, thus documenting the usefulness of situational self-report scores to decompose these sources of variance.

Finally, in terms of developing additional methods for testing validity, Chauliac et al. (2020) observed university students' gaze behavior while answering items on a questionnaire assessing habitual use of different cognitive strategies during learning from texts. There were systematic links between number and duration of fixations, on the one hand, and the consistency of answering different items from the same scale, on the other. The findings demonstrate that eyetracking has great potential to examine processes of responding to verbal stimuli as presented in self-report scales, suggesting that this methodology could contribute to examining the validity of scores derived from these scales.

Taken together, these six contributions attest to the potential validity of self-report in assessing students' learning. There were clear links (1) between quantitative self-report scores for different constructs, as well as (2) between these scores, on the one hand, and think-aloud protocols, online trace data, and academic performance, on the other. While not all of these links were fully robust and significant, they nevertheless document that self-report continues to be useful in measuring facets of students' learning.

4. Improving validity

How can we further improve the validity of self-report measures? Two of the contributions address this question. Fryer and Nakao (2020) examined the impact of type of response scale on levels, reliability, and factorial validity of self-reported task interest and its links with prior and subsequent domain interest in a sample of PhD students. Their study included two traditional formats (labelled categorical scale and visual analogue scale) as well as two more recent formats (slider and swipe scales). Reliability and factorial validity were nearly identical across the formats, and mean scores for interest in different tasks did not show systematic differences either. However, predictive validity for future interest tended to be higher for the slider and swipe versions than for the two traditional formats. This is promising and should stimulate research on how to further optimize response scales and their presentation.

Durik and Jenkins (2020) analyzed the link between undergraduate students' self-reported interest, their certainty in their answers, and their self-reported behavioral engagement in various subjects. The findings show that interest and certainty were related in a curvilinear fashion in most of the subjects; high certainty was associated with either low or high interest scores. Furthermore, the link between interest and behavioral engagement was substantially stronger for students with high certainty in their reported level of either individual or situational interest, and not even significant for students with low certainty in their situational interest. These findings suggest that including certainty ratings can increase the validity of self-report in predicting students' behavior. As such, although replication is needed, they represent a potential breakthrough in boosting the validity of interest measures.

In both of these contributions, it remains open to question how the observed effects can be explained. For the effects of certainty, as noted by Durik and Jenkins (2020), it seems possible that clear beliefs in the strength or weakness of one's interest contributes to using interest as a guide for action, in contrast to being unclear about one's interest which may subject action to situational conditions. Research on the origins and outcomes of certainty is needed to examine this possibility.

5. Directions for future research

5.1 Scoping a broader range of self-report methods

The contributions to this special issue focus on self-report methods using written verbal statements as stimuli and closed-format options to respond, with quantitative methods employed to analyze responses. However, as noted earlier, there are various alternative formats that are equally important. Each of these formats has its own advantages and disadvantages. Specifically, a substantial amount of social science research uses oral formats (specifically, interviews) and open-ended answers, either written or provided orally. To an extent, these formats may be subject to similar biases as written closed-format self-report, including response sets and memory biases. However, there also may be differences, especially in terms of strategies to reduce these problems.

For example, motivation to respond in socially desirable ways rather than telling the truth can be reduced by generating trust in interviewees that their data will be kept confidential, and memory biases can be reduced through cognitive interviewing techniques that optimize recall. Substantial progress in suitable methods has been made in forensic psychology and research on testimony (see, e.g., Bowles & Sharman, 2014; Brown & Lamb, 2015). It would be worth exploring if some of these strategies could be made fruitful for educational research as well. This may be especially important for research on learning in young children (preschool, kindergarten, and the early elementary school years).

Qualitative self-report methodology using open-ended answers is especially important when exploring new research questions, but also when wanting to understand unexpected or paradoxical findings that can be explored with in-depth interviews. How to best structure questions, analyze answers, and aggregate qualitative self-report findings across studies currently is a field of intense methodological debate (see, e.g., Clark, 2016; Snelson, 2016). Mainstream quantitative research on self-report should attend to these developments, and research is needed on how to better integrate different self-report methods and the resulting evidence (e.g., in terms of convergent parallel, exploratory sequential, or explanatory sequential mixed-method study designs; Creswell, 2014; Creswell & Plano Clark, 2011).

5.2 Importance of within-person research for validating self-report

Similar to other types of research in education and psychology, the vast majority of investigations using self-report have relied on between-person study designs, including most of the contributions to this special issue (the Dietrich et al., 2020, contribution is a notable exception). Between-person research is suited to examine individual differences and interindividual relations between variables. However, it is not suited to investigate the within-person psychological functioning that is addressed in theories of students' motivation, emotion, and self-regulated learning. Intraindividual and interindividual correlations are statistically independent, and there is no easy way to infer one from the other, except when conditions of ergodicity hold. These conditions include homogeneity of functional relations across persons and stationarity over time, conditions that are often not met (Voelkle et al., 2014). As such, to study motivation, emotion, and strategy use during learning, it is best to examine these processes within persons (Murayama et al., 2017). To ensure generalizability, the variation of within-person relations across persons needs to be analyzed – if there is little variation, then relations are generalizable and nomothetic conclusions can be reached.

The relevance of within-person research has important implications for the validation of self-report measures. In research on learning, some of these measures pertain to trait-like characteristics of students and are used to gauge between-person differences. For example, measures of trait-like individual interest may be used to assess differences in interest between students. For these measures, it is appropriate to use between-person designs to examine validity. However, whenever the purpose is to assess individual development over time, or personal functioning during learning, then it is more adequate to use within-person designs to validate self-report methods. Between-person designs can render misleading conclusions, and resulting findings can under- or overestimate validity relative to theories of individual learning.

5.3 The role of emotion and emotion regulation in self-report

As human performance more generally, adequately responding to self-report instruments requires both competence and motivation. Competence includes being able to understand questions, to retrieve relevant information from long-term memory or current working memory, and to integrate the retrieved information such that a decision about an adequate answer can be reached. Current models of self-report largely focus on these cognitive processes, and process-oriented methods to validate self-report items focus on techniques of cognitive validation (Castillo-Diaz & Padilla, 2013; Karabenick et al., 2007). Motivation includes wishes to veridically answer questions, either to get a valid self-assessment (e.g., in contexts such as career counselling or psychotherapy) or to help researchers in their attempts to understand reality, as well as desires to appear to others or oneself as a socially desirable person. Motivation has been examined especially in research on social desirability (see, e.g., Gignac, 2013), and there is a long-standing tradition of controlling for desirability in studies of personality.

However, beyond cognition and motivation, it seems likely that emotions also play a critically important role in self-report. Emotions are defined as affective responses to personally important events. As such, whenever self-report touches issues that are personally relevant, emotions are likely to be aroused. This can be emotions that are already associated with a given topic in memory, such as anxiety when retrieving recollections of prior exams, or emotions that are generated during the process of reading self-report items. In addition, emotions that are elicited by the task of responding and the social context of the assessment can play a role, such as sympathy for the experimenter administering a questionnaire, social anxiety of disclosing personal information, or anger about redundancy of items in lengthy multi-item instruments.

It is reasonable to assume that these emotions can substantially influence self-report responses. This can happen through the influence of emotions on retrieval of information from memory (e.g., in terms of mood-congruent retrieval), on integrating memory information in different ways (e.g., holistically in positive mood and detail-oriented in negative mood), on current motivation to persist in answering questions, and on motivation to answer in specific ways (e.g., according to social desirability when being socially anxious about one's responses). Furthermore, ways to regulate emotions may play a role as well. For example, unpleasant emotions triggered by emotionally negative self-report items may be so strong and aversive that one seeks to downregulate them right away, even before answering the item. As a result, the answer may no longer represent the original emotional response to the item. Emotions can contribute to changes of the objects of self-report measurement during the process of measuring them – a phenomenon that can render resulting scores an artefact of the response process.

Research exploring these possibilities is largely lacking. Self-report methodologists could team up with memory researchers, social psychologist, and affective scientists to investigate these possible influences of emotions on self-report. The results could inform psychological and educational measurement in terms of shaping instruments and the social situations of assessment in ways that are both emotionally beneficial and suited to increase validity.

6. Conclusion

Self-report is indispensable for any more fine-grained assessment of mental processes, including students' motivation, emotions, cognitive strategies, and metacognition during learning. Certainly, self-report has limitations in terms of assessing conscious processes only, being subject to biases, and not providing the temporal resolution needed to assess some of these processes. Nevertheless, the evidence reported in the contributions to this special issue clearly document that self-report continues to be a valid way to assess processes of learning. To further boost its validity, triangulation of different self-report methods (such as closed items and open-format think-aloud protocols) as well as integration of self-report into multi-channel assessments can be helpful. To make further progress in examining and improving the psychometric quality of self-report methods, it may be useful to consider a broad range of different variants of self-report, to consider the influence of respondents' emotions on their self-report, and to complement traditional between-person study designs with intraindividual analysis.

Keypoints

- Self-report is required for a nuanced assessment of mental processes
- Self-report is indispensable to assess learning-related emotions, motivation, meta-cognition, and self-regulation
- Learning-related self-report scales show convergent and predictive validity
- Self-report needs to be amended by alternative measures because it lacks temporal resolution and is subject to response sets and memory biases
- Future research should consider a broader range of self-report methods, within-person analysis, and the impact of emotions on self-report

References

Azevedo, R., Taub, M., & Mudrick, N.V. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. Schunk & Greene, J.A (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 254-270). Routledge.

Bowles, P. V., & Sharman, S. J. (2014). A review of the impact of different types of leading interview questions on child and adult witnesses with intellectual disabilities. *Psychiatry, Psychology and Law, 21*, 205–217. https://doi.org/10.1080/13218719.2013.803276

Brown, D. A., & Lamb, M. E. (2015). Can children be useful witnesses? It depends on how they are questioned. *Child Development Perspectives, 9*, 250-255. https://doi.org/10.1111/cdep.12142

Castillo-Diaz, M., Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research, 114*, 963–975. https://doi.org/10.1007%2Fs11205-012-0184-8

Chauliac, M, Catrysse, L. , Gijbels, D. & Donche V. (2020). It is all in the surv-eye: can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*, *8*(3), 26 – 39. https://doi.org/10.14786/flr.v8i3.489

Clark, A. M. (2016). Why qualitative research needs more and better systematic review. *International Journal of Qualitative Methods, 15*, 1-3. https://doi.org/10.1177/1609406916672741

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Sage.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage.

Durik, A. M. & Jenkins J. S. (2020). Variability in Certainty of Self-Reported Interest: Implications for Theory and Research. *Frontline Learning Research*, *8*(3) 85-103. https://doi.org/10.14786/flr.v8i3.491

Fontaine, J. J. R., Scherer, K. R., & Soriano, C. (Eds.). (2013). *Components of emotional meaning: A sourcebook*. Oxford University Press.

Fryer, L. K., & Nakao K. (2020). The Future of Survey Self-report: An experiment contrasting Likert, VAS, Slide, and Swipe touch interfaces. *Frontline Learning Research, 8*(3), 10-25. https://doi.org/10.14786/flr.v8i3.501

Gignac, G. E. (2013). Modeling the Balanced Inventory of Desirable Responding: Evidence in favor of a revised model of socially desirable responding. *Journal of Personality Assessment, 95*, 645–656. https://doi.org/10.1080/00223891.2013.816717

Iaconelli, R., & Wolters C.A. (2020). Insufficient Effort Responding in Surveys Assessing Self-Regulated Learning: Nuisance or Fatal Flaw? *Frontline Learning Research*, *8*(3) *104 – 125.* https://doi.org/10.14786/flr.v8i3.521

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Ree Bonney, C.,…& Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*, 139–151. https://doi.org/10.1080/00461520701416231

Lajoie, S. P., Pekrun, R., Azevedo, R., & Leighton, J. P. (in press). Understanding and measuring emotions in technology-rich learning environments. *Learning and Instruction*.

Moeller, J., Viljaranta, J., Kracke, B., & Dietrich, J. (2020). Disentangling objective characteristics of learning situations from subjective perceptions thereof, using an experience sampling method design. *Frontline Learning Research*, *8*(3), 63-84. http://doi.org/10.14786/flr.v8i3.529

Murayama, K., Goetz, T., Malmberg, L.-E., Pekrun, R., Tanaka, A., & Martin, A. J. (2017). Within-person analysis in educational psychology: Importance and illustrations. In D. W. Putwain & K. Smart (Eds.), *British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education – Current Trends: The Role of Competence Beliefs in Teaching and Learning* (pp. 71-87). Wiley.

Pekrun, R., & Bühner, M. (2014). Self-report measures of academic emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 561-579). Taylor & Francis.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*, 91-106. https://doi.org/10.1207/S15326985EP3702_4

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)* (Tech. Report No. 91-B-004). Board of Regents, University of Michigan, Ann Arbor, MI.

Rogiers, A., Merchie, E., & Van Keer, H. (2020). Opening the black box of students' text-learning processes: A process mining perspective. *Frontline Learning Research*, *8*(3) 40 – 62. https://doi.org/10.14786/flr.v8i3.527

Swann Jr, W. B., Chang-Schneider, C., & McClarty, K. L. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist, 62*, 84–94. https://doi.org/10.1037/0003-066X.62.2.84

Van Halem, N., van Klaveren, C., Drachsler H., Schmitz, M., & Cornelisz, I. (2020). Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data. *Frontline Learning Research*, *8*(3), 140-163. https://doi.org/10.14786/flr.v8i3.497

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Towards a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research, 49*, 193–213. https://doi.org/10.1080/00273171.2014.889593

Vriesema, C.C., & McCaslin, M. (2020) Experience and Meaning in Small-Group Contexts: Fusing Observational and Self-Report Data to Capture Self and Other Dynamics. *Frontline Learning Research*, *8*(3), 126-139. https://doi.org/10.14786/flr.v8i3.493