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.v8i3.625*

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