

Construct Validity of Self-Reported Metacognitive Learning Strategies

Authors and affiliation

Corresponding author:

Jean-Louis Berger

Swiss Federal Institute for Vocational Education

Avenue de Longemalle 1

1020 Renens

Switzerland

jean-louis.berger@iffp-suisse.ch

Phone: +41 21 621 82 31 or +41 78 760 33 53

Stuart A. Karabenick

University of Michigan

School of Education

Combined Program in Education and Psychology

610 East University Avenue

Ann Arbor, MI 48109-1259

USA

skaraben@umich.edu

Construct Validity of Self-Reported Metacognitive Learning Strategies

Abstract

Despite their significant contributions to research on self-regulated learning, those favoring online and trace approaches have questioned the use of self-report to assess learners' use of learning strategies. An important rejoinder to such criticisms consists of examining the validity of self-report items. Accordingly the present study was designed to establish the validity of ninth grade students' use of planning, monitoring, and regulation when studying math. To establish response process evidence of construct validity, cognitive interviews were coded to determine whether students' interpretations of the items' were consistent with their intended meaning and whether their response choices were congruent with those interpretations. Evidence supported the construct validity of monitoring and regulation items, but to a lesser degree those designed to assess planning. We discuss implications of the evidence for the self-report assessment of learners' use of metacognitive strategies.

Key words: Cognitive validity; Self-report; Questionnaire; Self-regulation; Metacognitive strategies

Introduction

According to Weinstein, Husman and Dierking (2000), learning strategies are “any thought, behaviors, beliefs, or emotions that facilitate the acquisition, understanding, or later transfer of new knowledge and skills” (p. 727). Metacognitive strategies are among those extensively studied. Given the essential role that strategies play in student learning, there have been concerted attempts to describe, foster, and train metacognitive strategies over the last three decades. How to measure the use of these strategies, which are by definition covert processes, however, remains controversial (Boekaerts & Corno, 2005; Dinsmore, Alexander, & Loughlin, 2008; Pintrich, Wolters, & Baxter, 2000; Veenman, 2005; Winne & Perry, 2000).

The most widely employed instruments to measure metacognitive and other learning strategies consist of self-report questionnaires with Likert-type response formats. A major advantage is that they can be completed quickly and easily by large numbers of students and are more cost-effective than are on-line, concurrent, methods. Moreover, self-report measures can be easily appropriated by school-psychologists, for example, for use as diagnostic tools. One review found that of 255 scrutinized empirical studies of these constructs, 43% used self-report either as an unique instrument or in combination with other instruments (Dinsmore et al., 2008). Examples are the Motivated Strategy for Learning Questionnaire (MSLQ; Pintrich et al., 1993) and Learning and Study Strategy Inventory (Weinstein, Palmer, & Shulte, 2002). Both rely on well-supported theoretical models of metacognition and self-regulated learning (SRL).

For some, however, self-report questionnaires have low construct validity (e.g., Tobias & Everson, 2000; Veenman, 2005; Winne et al., 2002; Winne & Perry, 2000), based on: a) cognitive limits of the respondent, b) the grain-size of assessment, and c) limited evidence for differential strategy use in students’ reports. These arguments are specific to the

Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment*, 21(1), 19-33.

topic of SRL and distinct from such well-known biases as social desirability, acquiescence and satisficing (Tourangeau et al., 2000).

Limits of Self-Reports

Cognitive limits of the respondent. Critics propose that learners' memory retrieval capacities may be insufficient to accurately report strategy use (Tobias & Everson, 2000), and rely instead, for example, on their familiarity with the strategy. Another claim is that if strategies are stored as automated scripts in long-term memory, and therefore operate without attention (Brown, 1987), learners may underestimate strategy use (Winne et al., 2002). Memory failure can appear at any stage of processing and responding to a self-report item: encoding, storage or retrieval (Tourangeau et al., 2000). Winne and Jamieson-Noel (2002) observed that while self-reports of strategy use in a specific task "unequivocally represent students interpretations about how they study, such self-reports may not accurately indicate what students actually do when they study" (p. 568). In a review of studies using multiple methods, Veenman (2005) assumed a categorical stance that there is "little or no correspondence between prospective and retrospective statements on the one hand and actual, concurrent behavior on the other" (p. 92), suggesting instead that concurrent measures (e.g., observation, thinking aloud) are more adequate to represent the use of metacognitive strategies than are questionnaires. Based on these claims, the reliance on self-reports would call into question approximately half of the studies of metacognition.

Grain-size. A second question concerns the appropriate level (school-general, domain-specific, task-specific or activity-specific) to investigate learning strategies. Concerning reports of learning strategies at a general level, several researchers (e.g., Winne et al., 2002) argue that this should be avoided because it has been empirically demonstrated that the context (domain-specificity) has a significant effect on (i.e., moderates) students' use of learning strategies (e.g., Hadwin, Winne, Stockley, Nesbit, & Woszczyna, 2001;

Vermetten, Lodewijks, & Vermunt, 1999;). Winne et al. (2002) also consider that not only are self-reported items too vague at a general level, even at the course level “there is ambiguity in what self-reports represent” (p. 137). Some items in the MSLQ, for example, are based on conditional if-then relations (e.g., “When I study for this class, I pull together information from different sources, such as lectures, readings, and discussions”). Even such conditionals, they argue, may refer to several different activities for a single course. Furthermore, according to Schellings (2011), it is even uncertain that the student follows the instructions to answer items with a particular course in mind.

Interdependence between the types of learning strategies. Another issue is whether strategies can be assessed independently. For example, whereas Pintrich et al. (1993) extracted (by confirmatory factor analysis) the expected number of factors from the MSLQ college version, the junior high school version of this same instrument did not allow the extraction (by exploratory factor analysis) of the same number of factors found in the college version. Most relevant here, metacognitive strategies were not separable from effort management strategies and thus were combined into a scale named “self-regulation” (Pintrich & De Groot, 1990). Furthermore, the three types of cognitive strategies (rehearsal, elaboration, and organization) formed a single scale and not the expected three scales. Similar issues have been observed in other instruments (Liu, 2009). Pintrich et al. (2000) concluded from the multiple studies revealing large interrelations between strategies that “students who engage in one component of self-regulated learning also engage in other components. Accordingly, efforts to separate the different components into theoretically smaller subcomponents may not be justified by the empirical data” (p. 81).

Advantages and Validity of Self-Reports

Yet others claim there is sufficient evidence to warrant the use of self-report measures (e.g., Wolters, Pintrich, & Karabenick, 2005; Weinstein et al., 2002). As concluded recently by Schellings and Van Hout-Wolters (2011):

Notwithstanding this criticism [of self-report instruments], we believe that the possibilities of large-scale testing and the practical usefulness should not be underestimated. Each method may have its own quality, but further research is needed to develop more precise measures. This type of research may not only be aimed at comparing different methods in one research design (i.e., multi-method research), but also at the characteristics of self-report instruments in order to improve the instruments or to construct alternative assessment measures with the advantages of self-reports. (pp. 85-86)

Accordingly, the present research takes the position that, rather than dismiss their legitimacy as some have suggested, given the advantages of self-reports in widespread and myriad applications, research is best served by improvements in their reliability and validity.

Cognitive Interviewing Approach to the Study of Self-Report Item Validity

Cognitive interviewing¹ is one approach to determining the degree to which self-report items are interpreted as intended, and thus the degree to which scale items contribute to evidence of construct validity, which is a variant of but not subsumed within validity based on response processes according to the Standards for Educational and Psychological Testing (2011), which we will refer to as item validity.² The information provided by the process can be used for item modification, as well as guidance concerning the construct that items are designed to operationalize. Cognitive interviewing is a more inclusive term that includes cognitive pretesting during scale development. The approach is based on an information-processing model and a procedure designed to acquire evidence to evaluate critical

components of that process (Karabenick, Woolley, Friedel, Ammon, Blazeovski, Bonney, De Groot, Musu, Gilbert, Kempler, & Kelly, 2007). Studies, such as Koskey et al. (2010), indicate that even if the psychometric qualities of a scale are adequate, items may not convey their intended meaning, and that small changes can alter what can be inferred from responses to those items.

Aim of the Present Study

Winne et al. (2002) noted that, at the time, no study of the cognitive processing of self-reported items—the objective of cognitive interviewing—to assess strategy use had been conducted. They believed, however, that the “methods by which learners create self-reports are themselves a topic worthy of study in research on SRL. Such studies into learners’ interpretations of current and past episodes of SRL offer interesting methodological puzzles of their own, puzzles that we predict are central to understanding more clearly what theorists should and can infer from self-report data” (p.151). In fact, self-report inventories are considered interventions designed “to cause the learner to recall or to generate a particular kind of response” (Winne & Perry, 2000, p.532), although it is important to stress that on-line measures, such as diary studies, are similarly obtrusive. Few formal applications of cognitive interviewing to items measuring learning strategies have been conducted (Clayton, Zusho, Barnett, Michna, & Hefter, 2008; Schellings, 2011). Schellings (2011) asked four 9th grade students to think aloud while answering a 58-item questionnaire assessing the use of metacognitive strategies when studying a history text. Her analyses showed that most items were properly understood. However, several terms—concepts, drawing conclusions, or finding information—were found to be confusing, which cautions against the use of abstract terms. Schellings also observed that students who found an item ambiguous tended to select a middle response alternative (anchored “sometimes”) that conveys a neutral answer, which was also observed when students were not able to recall their activities during studying.

Finally, incongruent answer choices were also observed, as when students chose a response that differed from their verbalized account of their study process. Unfortunately, Schellings did not report the frequency of those issues, and it is consequently unknown how severe the issues were. Clayton et al. (2008) examined the influence of culture (ethnicity) on the cognitive validity of MSLQ learning strategies items. Like Schellings, students found abstract terms such as “concept” ambiguous.

The present study addresses some of the self-report limitations highlighted above. First, regarding the grain-size issue, the cognitive interviews will indicate whether students refer to the appropriate topic when processing the item and whether there are confusions or a lack of precision due to the fact that the context provided in the items is too large or vague. Second, regarding the question of interdependence of learning strategies, the interviews will be designed to reveal potential mixes of strategies when students process the items; and whether students refer to a strategy other than the one targeted by the items. Finally, regarding the cognitive limits of the respondent, the interviews will be designed to focus on item meaning and interpretation rather than individual cognitive abilities.

Furthermore it has been noted that self-report instruments have generally been based on samples of successful students, who tend to use more strategies. However, there is no empirical evidence that the strategies used by those students will resonate with other, lower achieving, students. Since the resulting scales may not be valid for less successful students (Boekaerts & Corno, 2005), it was important to determine whether the understanding of items depends on the frequency of strategy use (Winne et al., 2002). Accordingly, two research questions addressed in the present study were: (a) How cognitively valid are items assessing metacognitive strategies? and (b) Does the item validity of items depend on the frequency of their use?

Definitions of Metacognitive Self-Regulation Components

A clear and operational definition of metacognitive strategies is a necessary condition to assess the validity of items that operationalize those constructs. Planning refers to activities performed before actually learning the material (i.e., forethought). These activities are generally covert and concern decisions about what has to be learned, and how it can or should be learned. Selection of a strategy, decisions about how much time and effort is required for its use, and setting a standard to be reached (i.e., its proximal goal) are prototypical planning activities (Brown, 1987; Zimmerman, 2008). Monitoring refers to activities performed either during or immediately after engaging in the learning process, and is generally considered an on-line process since it refers to the on-going activity. Monitoring activities are used to determine whether the standard that was set during the planning process has been reached; monitoring is also used to determine how close learners believe themselves to be to that standard. Prototypical monitoring activities are self-questioning, judgment of learning (i.e., subjective assessment of how well the material has been learned; Nelson & Narens, 1990), self-testing, and checking. Regulation refers to activities contingent on monitoring process' results (Pintrich, 2000a). Following assessment of the quality or quantity of learning, learners can either stop the learning process if they believe the standard has been reached, or regulate their learning activities. Prototypical regulation activities are returning to study the material again, slowing down to learn at a slower pace, trying a different way to learn (i.e., change the strategy), and persevering to learn the material. As the student learns, the activities are not necessarily performed in sequence. A student can skip the planning activity or choose to not regulate her/his learning processes. Furthermore, the activities can be schematically organized as a cyclical process (Zimmerman, 2000) going from planning to regulation and then back to planning.

Method

Participants

Participants were selected from a sample of 306 ninth graders enrolled in U.S. high school math classes, based on their total metacognitive self-regulation (MSR) score, which reflected the frequency of strategy use. In order to not only study the items' cognitive validity but also investigate whether this type of validity would differ based on the extent of reported use of MSR, 15 with the highest scores and the 15 students with the lowest scores were asked to participate in a 15 to 20 minute interview. Some students were not selected for the following reasons: (a) they were no longer enrolled in a math class, (b) they had dropped out of high school, (c) inability to complete the interview because of a severe personality disorder, (d) inability to even understand the interview questions or (e) data missing on the survey itself. Students eliminated for these reasons were replaced with those whose scores closely matched those originally selected. All of those finally selected accepted the invitation to participate. Table 1 describes the final sample. It should be noted that although students are not formally tracked in the school system, the students selected were distributed across five levels of math. The distribution of participants across the five levels of math courses for those high versus those low in MSR was not significantly different, $\chi^2(4, n = 29) = 1.14$, ns, nor was the gender distribution, $\chi^2(1, n = 29) = .54$, ns.

[Insert table 1 here]

MSR Scale Revision

MSR items used in the present study were based on metacognition scales of the college version of the MSLQ (Pintrich, et al., 1993). The psychometric properties of this instrument were described in several publications (Pintrich, et al., 1993; Pintrich et al., 1991; Garcia Duncan & McKeachie, 2005). As described below, several items were revised (see also Authors, 2011) to render them more appropriate for the subject matter (i.e., math) and

grade level. Further modifications were designed to improve their interpretability and to provide an acceptable level of discriminant validity between its three components (planning, monitoring, and regulation) given the metacognitive items used on the original MSLQ have been repeatedly found to be saturated by a single factor and thus combined into a single MSR scale (Pintrich & De Groot, 1990; Pintrich et al., 1993; Wolters et al., 2005) despite the items' face validity suggesting that the three components represented the distinct metacognitive strategies.

First, some items appeared to not assess MSR but rather attention control (e.g., *During class time I often miss important points because I'm thinking of other things; I often find that I have been reading for this class but don't know what it was all about*, both reversed items). Second, some items were long and/or were difficult to understand (e.g., *I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over when studying for this course*). Third, as Clayton et al. (2008) found, students varied in their interpretation of such words as “concept” (e.g., *When studying for this course I try to determine which concepts I don't understand well*). Fourth, one of the items (*I try to change the way I study in order to fit the course requirements and the instructor's teaching style*) included two aims (i.e., fit the course requirements and fit the instructor's teaching style) that rendered the respondents' referent ambiguous.

Planning items were modified to include references to “before” studying by adding the word “plan” to directly reference the temporal dimension of the construct that the item was designed to operationalize. Planning activities then included a selection of planning activities: what to study (goal setting), the selection of a learning strategy (how), and the time devoted in the study of a specific topic. In the monitoring items, the word “understand” was replaced by “know” or “have learned” in order to not confound the construct with achievement mastery goals that stress understanding and knowledge gains. In fact, including

the verb “to understand” was found by Koskey et al. (2010) to suggest both a deep learning strategy and the idea of striving for understanding. In the regulation items, the same conditional form was used for each item (“If...”) in order to more directly assess the construct. As regulation involves alteration of one’s approach to learning, items mentioned a context in which regulation would be appropriate. Contrary to the original MSLQ, the items concerned strategies specific to mathematics and not simply activities applied to one school domain (“In this class ...”). However, in order to be applicable to math in general, the items were not targeted to each math topic (e.g., algebra, geometry). Table 2 presents the original MSLQ items and the revised versions.

[Insert table 2 here]

The items are considered prototypical and sufficient to capture the critical aspects of each metacognitive strategy (i.e., the content aspect of construct validity; Messick, 1995). MSR items were first analyzed using the initial student sample to determine their psychometric characteristics (Authors, 2011). A model with three first order factors representing the three components of MSR and one second order factor saturated by each of the first order factors, shown in Figure 1 was found to adequately represent the data ($\chi^2_{\text{MLR}(64)} = 139.571, p < .001$; CFI = .91; RMSEA = .06), with the descriptive scale statistics presented in Table 3. The psychometric evidence thus supports the structural aspect of the scales’ construct validity (Messick, 1995), as well as their constituent items; the issue here is whether the same can be said for the items’ construct validity based on response process evidence.

[Insert figure 1 here]

[Insert table 3 here]

Cognitive Interview Procedure

Cognitive interviews were conducted individually in a separate room during the students’ regular math classes. Students were informed that the researchers wanted to know

what some of the questions from the survey meant to them and what they thought when answering them. We stressed the fact that it was important for them to know that there were no right or wrong answers to the questions, only what they thought was right for them: “We really value what you personally think about the questions as well as why you pick a certain number”. The students were free to skip any questions they do not feel comfortable answering, and stop the interview at any time.

The standardized interview protocol specified by Karabenick et al. (2007) and Koskey et al. (2010) was identical for each item (see Appendix A). Students were first asked to read the item aloud, then tell what they thought the item was asking, to select an alternative on the 1 to 5 scale (from Not at all true to Completely true) and then to explain why she chose that answer. Standardized probes followed to obtain information regarding several aspects of item validity. First, item *interpretation* was judged based on answers to the probe “What is that question trying to find out from you?” Second, what is termed the student’s coherent *elaboration* was based on responses to the probe “Please explain to me why you chose that answer? What were you thinking about?” Finally, student’s *congruence of answer choice* was judged based on the comparison between the selected answer and the rationale given for that answer. Follow-up probes and requests to reread the item were used, respectively, to elicit additional information from students giving too little information and to make the student aware of a mistake in when reading the item. Interviews were audio-recorded and transcribed. Items were presented in random order.

Coding System Development and Coding

Analogous to Koskey et al. (2010) and Wooley et al. (2006), a coding system was developed based on the metacognitive construct (i.e., planning, monitoring, regulation) it was designed to operationalize: a) how each item should be interpreted, b) the elaboration that would be appropriate for the item, and c) the answer choice that would be congruent with

students' elaboration. An example of the coding system for one of the items (No. 4) is presented in Appendix B. In this example, item interpretation was rated as congruent or not (dichotomous rating) by comparing student's interpretation with that expected based on the construct being operationalized. In the example shown, the response was specified as congruent if "The student thinks about what would be the most efficient way to study something new. The 'best way' can refer to a strategy to learn better (e.g., understand more deeply, memorize) or to a less time-consuming strategy." It was rated as incongruent if the student's interpretation did not match this description. An additional requirement was that the student did not refer to a topic other than math. Elaboration was subdivided in several ratings depending on the item content. Rating A (when a moment or a condition) referred to the time reference (before, during/right after studying) or, for the items assessing regulation, the condition of application (e.g., *if I get confused...*). Rating B (what) referred to the kind of activity or strategy. Rating C (why) referred to the aim of the activity, which was applied to only three items explicitly mentioning an aim. Some items included two A ratings since two conditions or moments were mentioned (e.g., "I plan how I am going to study new math topics before I begin" refers to an activity undertaken before actually beginning to study and in relation to a new topic). These criteria will become more evident when examples are provided subsequently.

Results

Interrater Reliability

Interrater was estimated by agreement between one of the authors and a second rater who coded half of the items. Both raters coded all students on an item before moving to the next item. For each item, students' transcripts were coded in random order, and coders were blind to the students' total MSR scores. Interrater reliability was computed, using Cohen's kappa, for six of the thirteen items (total of ratings = 26 ratings x 29 students = 754). The

mean kappas were: $\kappa = .76$ (interpretation; ranging from .50 to 1.00), $\kappa = .80$ (elaboration; ranging from .51 to 1.00), and $\kappa = .95$ (answer choice; ranging from .77 to 1.00), which collectively provide evidence of satisfactory interrater reliability.

Evidence of Item Validity

The percentages of coherent interpretation, congruent elaboration, and congruent answer choice are shown in Table 4. Planning items are the most problematic in terms of item validity as four out of five items resulted in low percentages ($<66\%^3$) for one or more aspects of item validity. By contrast, monitoring and regulation items present only moderate issues in interpretation, and there were no problems for the other item validity criteria. The congruence of answer choice was not problematic for the students as it was at least congruent in 86% of the cases. Moreover, referring to another topic in school instead of math was very infrequent. The following sections present results for each MSR component consecutively, providing interview excerpts to illustrate the difficulties students encountered when thinking about the item and elaborating their answers. The emphasis is placed on planning items given that they exhibited more validity problems than did the monitoring and regulation items.

[Insert table 4 here]

Validity of Planning Items. Two planning items (No. 2: *Before I begin studying math I think about what and how I am going to learn*; No. 4: *When I learn new topics in math, I first figure out the best way to study*) were coherently interpreted by less than half of the students (respectively 41% and 35%), whereas the other three items were properly interpreted by at least three fourth of the students. Taking into account the other item validity criteria, No. 4 was the poorest performing item (*When I learn new topics in math, I first figure out the best way to study*). For this item, only 35% of the students responded with cognitively valid elaborations, and this item showed the lowest percentage of congruent answer choices (86%). The problems in interpreting this item appear primarily due to students skipping the idea of

“new topics” and mentioning no learning tasks or other tasks such as test preparation, which does not fit the item meaning as defined a priori. Accordingly, students elaborated and then provided a rating based on different conditions than those stated in the item. Regarding elaboration, several students did not refer to an activity before beginning to study but rather one engaged in at any moment during the learning process. Although thinking about the best way to study is profitable not only before beginning to learn but also while learning, the item’s time reference was judged incongruent in these cases.

Among the general difficulties students encountered when interviewed about the planning items was the lack of metacognitive knowledge about the strategies. Hence, some students were not able to describe the learning strategies; they were not familiar with them and did not have the necessary vocabulary or knowledge in order to discuss them. In addition to the issue of metacognitive knowledge, this student mixed the planning strategy with other strategies in her interpretation. In fact, her interpretation “sitting down at somewhere quiet” refers to what is called “time and study environment management” in the MSLQ. Hence, her answer was not related to the expected construct. The mix of learning strategies constructs is also exemplified in the following excerpt (same item) where the student confounded planning and help seeking, together with a misunderstanding of the time reference. Among the five items designed to tap planning strategies, only one (No. 3: “Before I study math, I plan how much time I will need to learn a topic”) presented adequate item validity in our sample, presumably because it is a more concrete activity, easier for students to represent and verbalize.

Validity of Monitoring Items. Monitoring items demonstrated adequate item validity in all the aspects considered, with no validity category that was under 66%. Nevertheless, a recurrent issue concerned the fact that students sometimes mentioned regulation activities whereas the item asked about monitoring strategies. This lack of congruent elaboration was

repeatedly found, and it may provide an explanation of why these two components are so strongly related in students' self-reported levels of these activities (Pintrich et al., 2000).

Item validity of Regulation Items. Although the interpretation of these items was not as problematic as the interpretation of planning items, two items (Nos. 11 & 12) were incorrectly interpreted by 31% of the sample. The main reason is that students narrowed the meaning of the items or skipped part of the item when interpreting it. Paralleling the observation that students talk about regulation activity when elaborating or interpreting monitoring items, they also reported monitoring in the context of regulation items.

Differences Between the Item Validity of High and Low MSR Groups

MANOVA was employed to determine whether high and low MSR groups differed in their level of item validity. The MSR group was the fixed factor and the dependent variables were the mean scores for each item validity criterion. The results revealed a significant effect of group: Pillai's Trace, $F(6,22) = 3.12, p = .02, \eta^2 = .46$. ANOVAs, as shown in Table 5, revealed that the scores in two item validity criteria were significantly lower for low-scorers compared to high scorers: Interpretation ($F(1,28) = 5.94, p = .02, \eta^2 = .18$) and Elaboration A2 (second reference to time or condition of application) ($F(1,28) = 10.42, p = .003, \eta^2 = .28$).

[Insert table 5 here]

Hence, the students scoring the lowest tend to provide less adequate interpretation of the items and to elaborate less coherently than did students scoring the highest. Further analysis using the mean scores again for the validity criteria for each scale independently, indicated that the above-mentioned differences were manifest in the planning items but not in the monitoring and regulation items. We can therefore conclude that the level of item validity for planning items was lower for students lower on their overall MSP scores.

Discussion

The present study examined the item validity of self-report items designed to assess metacognitive strategy use by ninth-grade students in their math classes, using a modified version of the college form of the MSLQ (Pintrich et al., 1993). Second, whether the item validity of items depended on the frequency of their use. Psychometric analyses revealed that planning, monitoring, and regulation scales had acceptable psychometric properties. Cognitive interviews showed, in addition, that items used to assess monitoring and regulation have adequate evidence of validity. However, we found that, despite item improvements, the “classical” planning items indicated weak item validity. Accordingly, the scores based on those items lack the degree of construct validity compared to those used to assess monitoring and regulating, even though the psychometrics of the latter were not markedly better. Specifically, two planning items were not understood according to researchers’ expectations by more than half of the students, raising important concern about their validity. Only one item in the planning scale demonstrated satisfactory item validity on all the aspects scrutinized. Our analysis, therefore, suggests not using the planning scale in its present form with high school or younger students in math, and potentially in other subjects as well.

Students also had different interpretations of the planning items depending on their reported use of these strategies. Those reported having used planning strategies the most had a more coherent interpretation of the items than did those not planning as frequently. Metacognitive knowledge about (or awareness of) one’s own strategies may also be a factor explaining difference between low- and high-scorers. Students who developed a detailed image of how they study, which strategies are appropriate in which context, and about the utility and cost of strategies may have been more prone to think about their responses to items

such as those in the present study (Boekaerts & Corno, 2005; Winne et al., 2002), and were more likely to verbalize such activity.

Evidence thus suggests that items to assess planning strategies in mathematics are in need of further revision. This includes using item content more concrete and that targets simpler planning activities. Such revisions should take into account students' difficulties in thinking about planning activities. In metacognitive theories, planning is assumed to be a conscious activity, or at least considered to undertake with higher levels of awareness than are other metacognitive strategies, such as monitoring (Brown, 1987). Results in the present study suggest however, that ninth graders may lack awareness of the planning strategies they use (Winne et al., 2002). Furthermore, even if students know how to plan, they may rely primarily on teachers' and parents' directions to study, rather than planning their work themselves (Zimmerman & Martinez-Pons, 1986).

In addition to the low item validity of planning items, most striking was interview content that revealed students mixing conceptually different learning strategies (especially monitoring and regulation) when interpreting and elaborating their answers. This may provide part of the explanation for why learning strategies are so strongly correlated in metacognition studies of high school students. Results from the CFA we conducted corroborate this result, as all three metacognitive scales loaded on a second-order metacognition factor. A complementary explanation might be found in the interdependence of monitoring and control processes, as formulated in Nelson and Narens' (1990) model of metacognition: The two types of *metacognitive processes* are conceived as flows of information between the object-level and the meta-level constituting a classical model of metacognition. Accordingly, a statistical correlation and intermingling of metacognitive strategies is to be expected. Furthermore, students mixed self-regulation with other-regulation, especially when indicating that they sought help from their teacher. Whereas the

items were designed to assess self-regulation of cognition, interviews revealed that students also thought about other- or co-regulation when interpreting or elaborating their answers. An additional issue was that students just ignored strategy aim (e.g., “to help me learn”) when included in item wording. It also seems clear that all the strategies aimed at facilitating or improving learning. Therefore, we conclude that such broad strategy aims should be replaced by more specific aims related to outcomes. An instance of specific outcomes for using strategies is “to get high grades”, “to succeed in this class”. Note that specifying aims might provide additional confusion with motivational constructs. In fact, adding the outcome of understanding or getting the best grades of the class refers to mastery and performance achievement goals and thus increase construct-irrelevant variance in the strategy we strive to assess.

Students demonstrated that their answer choices were largely congruent with their thoughts about using the strategies, in contrast to observations by Schellings (2011) based on ninth graders. This may be due to our structured interview approach that systematically asked students to provide a rationale for their answer, and directed them to respond to items keeping in mind the domain of mathematics. It seems that a more important issue, observed in particular in planning items, is that students sometimes referred to activities different from those mentioned in the item, which effectively altered the item meaning. Our use of a precise coding system may also have contributed to more accurately determining the level of response congruence.

It should be emphasized that the present study does not allow generalization of the results to older students. Obviously, the ability to introspect is still developing at the age of 15, and thus the items may well evidence higher item validity for older students. A comparison across age groups could provide essential information regarding differences in item validity, as it would for any other source of validity, including construct validity in

general. Assuming that college students are more familiar with planning (a claim some consider questionable), items designed to detect planning should manifest higher levels of item validity. College students' cognitive abilities should also allow them to understand items that include terms with higher levels of abstraction.

One may also wonder how the results would generalize to other self-reported measures of metacognition or self-regulated learning and also, more broadly to self-reported scales targeting other constructs. Based on our knowledge of other self-report instruments in the field, similar weaknesses, notably an overlap of multiple strategies in student thinking leading to mixing conceptually different learning strategies, may be present. For sure, specific cognitive validity studies are called for to improve the items of those instruments, keeping in mind that result may vary depending on population of interest. In the footsteps of studies by others (Karabenick et al., 2007; Koskey et al., 2010; Wooley et al., 2006), the present study constitutes an additional example of the value of cognitive interviews in complementing the psychometric perspective on item validity. Even though some of the issues revealed likely apply to other scales (e.g., the difficulty in understanding highly abstract items; also revealed in other cognitive interview studies by Kosey et al. and Wooley et al.), each construct would also have specificities that lead us to conclude that those scales warrant specific cognitive interview studies.

Further research may investigate additional solutions to improve validity of information provided by self-report questionnaires. For instance, would more detailed task-specific items on planning strategies (e.g., trying to solve simultaneous mathematical equations) show higher item validity than subject-specific items (e.g., in math, or even in algebra)? The hypothesis here would be that more focused and precise items reduce room for multiple interpretations and hence are interpreted more closely to researchers' expectations than subject-specific items. In other words, this approach may reduce construct-irrelevant variance

as Schellings' study suggests (Schellings, 2011). The cognitive validation approach taken here could provide crucial evidence beyond that generally adduced to provide answers to such questions.

References

- Berger, J.-L., & Karabenick, S. A. (2011). Motivation and students' use of learning strategies: Evidence of unidirectional influences in mathematics classrooms. *Learning and Instruction, 21*(4), 416-428. doi: 10.1016/j.learninstruc.2010.06.002
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An international Review, 54*(2), 199-231. [doi:10.1111/j.1464-0597.2005.00205.x](https://doi.org/10.1111/j.1464-0597.2005.00205.x)
- Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Erlbaum.
- Clayton, K. E., Zusho, A., Barnett, P. A., Michna, G., & Hefter, S. (2008, March). *Indices of achievement goals and metacognitive strategy use: Are they culturally and cognitively valid?* Paper presented at the Conference of the American Educational Research Association, New York, March, 24-28.
- 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. [doi:10.1007/s10648-008-9083-6](https://doi.org/10.1007/s10648-008-9083-6)
- Garcia Duncan, T., & McKeachie, W. J. (2005). The making of the Motivated Strategies for Learning Questionnaire. *Educational Psychologist, 40*(2), 117-128. doi:10.1207/s15326985ep4002_6
- Hadwin, A. F., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderate students' self-reports about how they study. *Journal of Educational Psychology, 93*(3), 477-487. [doi:10.1037/0022-0663.93.3.477](https://doi.org/10.1037/0022-0663.93.3.477)
- Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., De Groot, E., Musu, L., Gilbert, M. C., Kempler, T. M., & Kelly, K. L. (2007). Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment, 21*(1), 19-33.

Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, 42(3), 139–151.

[doi:10.1080/00461520701416231](https://doi.org/10.1080/00461520701416231)

Koskey, K. L. K., Karabenick, S. A., Wooley, M. E., Bonney, C. R., & Dever, B. V. (2010).

Cognitive processing in students' judgments of classroom mastery goal structure:

What are they thinking and why it matters. *Contemporary Educational Psychology*,

35(4), 254-263. [doi:10.1016/j.cedpsych.2010.05.004](https://doi.org/10.1016/j.cedpsych.2010.05.004)

Liu, O. L. (2009). Evaluation of a learning strategies scale for middle school students.

Journal of Psychoeducational Assessment, 27(4), 312-322.

[doi:10.1177/0734282908327935](https://doi.org/10.1177/0734282908327935)

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), 741-749. [doi:10.1037/0003-066X.50.9.741](https://doi.org/10.1037/0003-066X.50.9.741)

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning

components of classroom academic performance. *Journal of Educational Psychology*,

82(1), 33-40. [doi:10.1037/0022-0663.82.1.33](https://doi.org/10.1037/0022-0663.82.1.33)

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategy for Learning Questionnaire (MSLQ)*. University of

Michigan: NCRIPTAL.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and

predictive validity of the motivated strategies for learning questionnaire (MSLQ).

Educational and Psychological Measurement, 53, 801-813.

[doi:10.1177/0013164493053003024](https://doi.org/10.1177/0013164493053003024)

Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment*, 21(1), 19-33.

- Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43-97). Lincoln, NE: Buros Institute.
- Schellings, G. (2011). Applying learning strategy questionnaires: Problems and possibilities. *Metacognition and Learning*, 6, 91-109. doi: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, 83-90. doi:10.1007/s11409-011-9081-9
- American Educational Research Association (2011). *Standards for Educational and Psychological Testing*. Washington, DC: Author.
- Tobias, S., & Everson, H. T. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 147-222). Lincoln, NE: Buros Institute.
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. New York: Cambridge University Press.
- Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artlet & B. Moschner (Eds.), *Lernstrategien und metakognition: Implikationen für forschung und praxis* (pp. 77-99). Berlin: Waxmann.
- Vermetten, Y. J., Lodewijks, H. G., & Vermunt, J. D. (1999). Consistency and variability of learning strategies in different university courses. *Higher Education*, 37(1), 1-21. doi:10.1023/A:1003573727713
- Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Self-regulation interventions with a focus on learning strategies. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 727-747). San Diego, CA: Academic Press.
- Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment*, 21(1), 19-33.

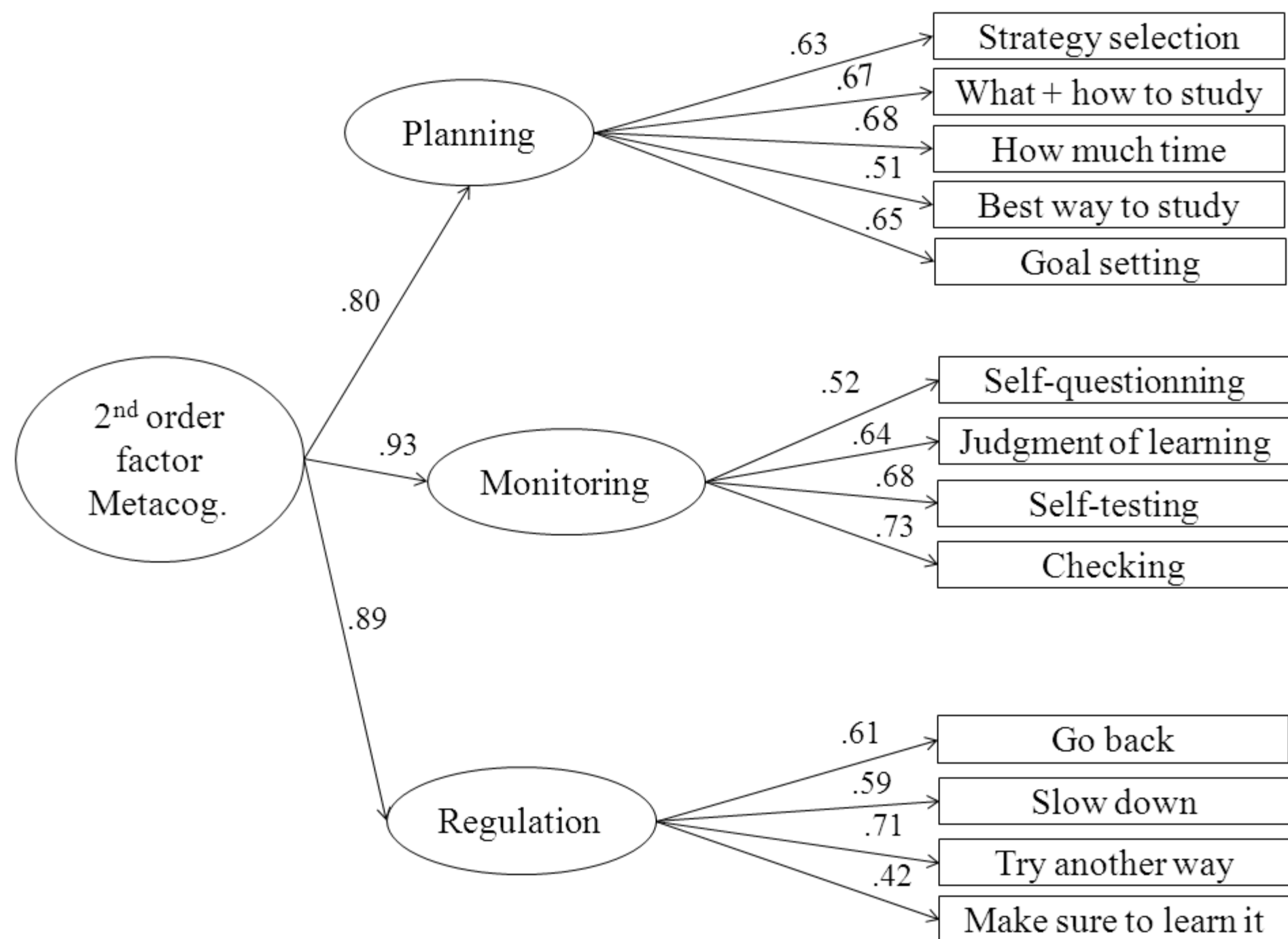
- Weinstein, C. E., Palmer, D. R., & Shulte, A. C. (2002). *Learning and Study Strategies Inventory* (2nd ed.). Clearwater, FL: H & H Publishing Company.
- Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology*, 27, 551-572. [doi:10.1016/S0361-476X\(02\)00006-1](https://doi.org/10.1016/S0361-476X(02)00006-1)
- Winne, P. H., Jamieson-Noel, D., & Muis, K. R. (2002). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in Motivation and Achievement* (Vol. 12, New directions in measures and methods, pp. 121-155). Amsterdam: JAI.
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531-566). San Diego, CA: Academic Press.
- Wolters, C. A., Pintrich, P. R., & Karabenick, S. A. (2005). Measuring academic self-regulated learning. In K. A. Moore & L. Lippman (Eds.), *Conceptualizing and measuring indicators of positive development: What do children need to flourish?* (pp. 251-270), New York: Kluwer
- Wooley, M. E., Bowen, G. L., & Bowen, N. K. (2006). The development and evaluation of procedures to assess child self-report item validity. *Educational and Psychological Measurement*, 66(4), 687-700.
- Zimmerman, B. J. (2000). Attaining self-regulation. A social cognitive perspective. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego, CA: Academic Press.
- Zimmerman, B. J., & Martinez Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23(4), 614-628. [doi:10.3102/00028312023004614](https://doi.org/10.3102/00028312023004614)
- Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment*, 21(1), 19-33.

Footnotes

1. The term cognitive interviewing is used in preference to cognitive pretesting that applies more specifically to the development phase of scale construction.
2. The term item validity has been used to designate the study of respondents' cognitive processes as they answer self-report items (Karabenick et al., 2007; Koskey et al., 2010) or in the domain of performance assessment in math and science. Karabenick et al. (2007) used the term item validity to refer to the contribution of information obtained using cognitive interviewing and related techniques, such as think aloud procedures, to construct validity. Most generally, item validity is not intended to introduce another validity category but should be viewed as similar to providing "evidence based on response processes" according to the Standards for Educational and Psychological Testing (2011).
3. An item congruently interpreted by less than 66% of the students was considered problematic.

Figure Caption

Figure 1. Confirmatory factor analysis model of the MSR items ($n = 306$).



Appendix A: Interview protocol

1. “Please read that question out loud for me.”

If the student has trouble:

- a. Read aloud to the student the word(s) he or she had trouble with.
- b. Ask the student if he or she understands the word(s) after hearing you say it.

2. “What is that question trying to find out from you?”

If the student repeats exactly the item:

Is there any other way to say it? or Tell me this in your own words please

3. “These numbers describe how different people feel about this (question/idea). Which number would you choose as your answer?”

4. “Please explain to me why you chose that answer. What were you thinking about?”

If the student does not report memories about using the strategy:

- a) Can you tell me a little more about why you chose number ____?
- b) Can you give me an example?
- c) Are there any other reasons why you chose number ____?

5. “Now let’s do the next question.”

Appendix B Sample Coding System

Item 4

“When I learn new topics in math, I first figure out the best way to study.”

Interpretation

Valid if the student thinks about what would be the most efficient way to study something new. The “best way” can refer to a strategy to learn better (understand more deeply, memorize it) or to a less time-consuming strategy.

Elaboration

Valid if...

A. Student refers to an activity undertaken:

- before beginning *and*
- For new topic (both have to be present)

B. Student has to mention a choice of a way to study (e.g., doing the exercises, reading the theory/notes first, memorizing)

C. Not valid if no aim mentioned

Congruent Answer choice (score)

5 or 4 if student reports often thinking about the best way to study

3 if student reports sometimes thinking about the best way to study

2 or 1 if student reports never or not knowing how to think about the best way to study

Table 1.

Sample Description by Group

| MSR Group | <i>n</i> | %Female | MSR Scores | | |
|-----------|----------|---------|------------|-----------|---------|
| | | | <i>M</i> | <i>SD</i> | Range |
| Low | 15 | 40 | 19.73 | 3.86 | 15 - 27 |
| High | 14 | 64 | 50.47 | 2.36 | 45 - 55 |
| Total | 29 | 52 | 35.10 | 15.94 | 15 - 55 |

Note. MSR = Metacognitive Self-regulation.

Table 2.

Original and Revised MSLQ items

| Original MSLQ items (college students) | Revised items adapted to math (high school students) |
|---|---|
| Goal Setting or Planning | |
| When reading for this course, I make up questions to help focus my reading. | 1. I plan how I am going to study new math topics before I begin. |
| When I study for this class, I set goals for myself in order to direct my activities in each study period. | 2. Before I begin studying math I think about what and how I am going to learn. |
| I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over when studying for this course. | 3. Before I study math, I plan how much time I will need to learn a topic. |
| Before I study new course material thoroughly, I often skim it to see how it is organized. | 4. When I learn new topics in math, I first figure out the best way to study. |
| | 5. Before I study math, I set goals for myself to help me learn. |
| Comprehension Checking (Monitoring) | |
| When I become confused about something I'm reading for this class, I go back and try to figure it out. | 6. When I study math, I ask myself questions to make sure I know what I have been learning. |
| I ask myself questions to make sure I understand the material I have been studying in this class. | 7. When studying math I try to determine how well I have learned what I need to know. |
| I often find that I have been reading for this class but don't know what it was all about. | 8. When I'm studying math I test myself to see whether I know the material. |
| When studying for this course I try to determine which concepts I don't understand well. | 9. I check whether I have learned what I am studying in math. |
| Regulation | |
| If course readings are difficult to understand, I change the way I read the material. | 10. If I get confused with something I'm studying in math, I go back and try to figure it out. |
| I try to change the way I study in order to fit the course requirements and the instructor's teaching style. | 11. If the math I am studying is difficult to learn, I slow down and take my time. |
| If I get confused taking notes in class, I make sure I sort it out afterwards. | 12. If I'm having trouble solving math problems I try other ways to solve them. |
| During class time I often miss important points because I'm thinking of other things. | 13. If I think I don't know my math well enough, I make sure I learn it before going to the next topic. |

Table 3.

Descriptive Statistics and Internal Consistency of MSR Component Based on CFA

| Component Scale | # items | M | SD | Skew | α |
|-------------------------------------|---------|------|-----|-------|----------|
| Planning | 5 | 2.10 | .82 | .625 | .75 |
| Monitoring | 4 | 3.23 | .99 | -.343 | .78 |
| Regulation | 4 | 3.73 | .87 | -.809 | .66 |
| 2 nd order Metacognition | 13 | 3.03 | .74 | -.292 | .86 |

Note. α = Cronbach's alpha estimate of internal consistency.

Table 4.
Percentage of Students Meeting Congruent Interpretation, Coherent Elaboration, and Congruent Answer choice Criteria

| Item | Interpretation | Elaboration A1 | Elaboration A2 | Elaboration B1 | Elaboration B2 | Elaboration C | Answer Choice |
|---|----------------|-------------------|-------------------|-------------------|-------------------|------------------|------------------|
| Planning | | | | | | | |
| 1. I plan how I am going to study new math topics before I begin. | 76 | 97 | 48 | 62 | NA | na | 90 |
| 2. Before I begin studying math I think about what and how I am going to learn. | 41 | 93 | na | 72 | 69 | na | 100 |
| 3. Before I study math, I plan how much time I will need to learn a topic. | 76 | 79 | na | 79 | na | na | 93 |
| 4. When I learn new topics in math, I first figure out the best way to study. | 35 | 62 | 45 | 93 | na | na | 86 |
| 5. Before I study math, I set goals for myself to help me learn. | 86 | 72 | na | 86 | na | 52 | 100 |
| Monitoring | | | | | | | |
| 6. When I study math, I ask myself questions to make sure I know what I have been learning. | 83 | 72 | na | 93 | na | 69 | 100 |
| 7. When studying math I try to determine how well I have learned what I need to know. | 83 | 83 | na | 79 | na | na | 90 |
| 8. When I'm studying math I test myself to see whether I know the material. | 72 | 79 | na | 93 | na | 76 | 93 |
| 9. I check whether I have learned what I am studying in math. | 82 | 100 | na | 86 | na | na | 97 |
| Regulation | | | | | | | |
| 10. If I get confused with something I'm studying in math, I go back and try to figure it out. | 90 | 97 | na | 97 | na | na | 97 |
| 11. If the math I am studying is difficult to learn, I slow down and take my time. | 69 | 79 | na | 93 | na | na | 97 |
| 12. If I'm having trouble solving math problems I try other ways to solve them. | 69 | 83 | na | 90 | na | na | 100 |
| 13. If I think I don't know my math well enough, I make sure I learn it before going to the next topic. | 83 | 86 | 93 | 83 | na | na | 93 |

Note: Elaboration A refers to the strategy timing; Elaboration B refers to the strategy itself; Elaboration C refers to the strategy aim.

Table 5.
Cognitive Validity Criterion Scores for Low and High MSR Groups

| Validity Criterion | MSR Group | M | SD | Range | F | <i>p</i> | η^2_p |
|---|-----------|-----|-----|------------|-------|----------|------------|
| Interpretation | Low | .68 | .11 | .46 – .85 | 5.94 | < .05 | .18 |
| | High | .80 | .14 | .54 – 1.00 | | | |
| | Total | .74 | .14 | .46 – 1.00 | | | |
| Elaboration A1 (time reference or condition of application) | Low | .84 | .15 | .54 – 1.00 | < 1 | ns | - |
| | High | .85 | .13 | .62 – 1.00 | | | |
| | Total | .84 | .14 | .54 – 1.00 | | | |
| Elaboration A2 (time reference or condition of application) | Low | .49 | .21 | .00 – .67 | 10.42 | < .01 | .28 |
| | High | .76 | .24 | .33 – 1.00 | | | |
| | Total | .62 | .26 | .00 – 1.00 | | | |
| Elaboration B1 (type of activity/strategy) | Low | .85 | .13 | .58 – 1.00 | < 1 | ns | - |
| | High | .90 | .14 | .62 – 1.00 | | | |
| | Total | .87 | .13 | .58 – 1.00 | | | |
| Elaboration B2 (type of activity/strategy) (single item) | Low | .67 | .49 | .00 – 1.00 | < 1 | ns | - |
| | High | .71 | .47 | .00 – 1.00 | | | |
| | Total | .69 | .41 | .00 – 1.00 | | | |
| Elaboration C (aim of the activity) | Low | .58 | .34 | .00 – 1.00 | 2.34 | ns | - |
| | High | .74 | .19 | .33 – 1.00 | | | |
| | Total | .66 | .29 | .00 – 1.00 | | | |
| Answer choice | Low | .96 | .06 | .85 – 1.00 | < 1 | ns | - |
| | High | .95 | .07 | .77 – 1.00 | | | |
| | Total | .95 | .06 | .77 – 1.00 | | | |