Metacognition Learning
DOI 10.1007/s11409-017-9176-z

Evaluating the metacognitive awareness inventory using empirical factor-structure evidence

George M. Harrison[1] · **Lisa M. Vallin**[1]

Received: 23 November 2016 / Accepted: 18 July 2017

Abstract Many scholars agree on the general theoretical structure of metacognition, which is what informed the development of the Metacognitive Awareness Inventory (MAI). Although self-report instruments such as the MAI suffer many threats to validity, they continue to be used in research and practice because of their convenience. With the MAI, studies have varied in the way they calculate scores and in their adherence to the intended theory. In this study, we address these shortcomings and propose modifications in calculating MAI scores. Using confirmatory factor analysis (CFA) and multidimensional random coefficients multinomial logit (MRCML) item-response modeling, we examined how well the intended functioning of the MAI matched the data from 622 undergraduate students. The results support scoring the MAI as two dimensions, knowledge and regulation of cognition, but indicate that the 52-item instrument has poor fit. Using iterative CFA and MRCML models, we tested subsets of items that represent the theory and had good fit. We followed up with tests of between-group and time invariance. The results support the use of a 19-item subset for between-group comparisons, with provisional evidence for its use in longitudinal studies.

Keywords Assessment of metacognition · Self-report · Confirmatory factor analysis · Item-response modeling · Invariance

The emphasis on metacognition exists in several corners of education research. In collaborative learning research, it constitutes a component of collective learning of content (Khosa and Volet 2014; Vauras et al. 2003). In motivation research, it constitutes a component of successful learning strategies (e.g., Duncan and McKeachie 2005). In the critical thinking literature, it is often considered a foundation to critical thinking (Halpern 1998; Ku and Ho 2010; Kuhn and Dean 2004; Magno 2010; Schön 1983).

✉ George M. Harrison
georgeha@hawaii.edu

[1] Curriculum Research & Development Group, College of Education, University of Hawai'i at Mānoa, Honolulu, HI, USA

Published online: 24 July 2017

Though an agreed-upon definition of metacognition has been elusive (Dinsmore et al. 2008), a longstanding view is that it comprises two components: knowledge of cognition and regulation of cognition (Brown 1987), where the former involves our awareness of our thought processes, particularly our declarative knowledge about our memory (Flavell 1979), and the latter involves our planning and control of these processes (Brown 1987; Jacobs and Paris 1987). Being able to control our cognition, and therefore apply metacognitive skills, requires knowledge of various strategies and awareness about when to best apply them (Ambrose et al. 2010; Schraw et al. 2006). An individual's regulation of cognition involves a continuous evaluation of what is known and what still needs to be learned (Brown 1987; Flavell 1976; Jacobs and Paris 1987).

Research on metacognition has increased in frequency since the 1970s. Practitioner oriented studies have investigated correlations between metacognitive awareness and achievement (e.g., RincónGallardo 2009; Young and Fry 2008), with some studies presenting evidence or cogent arguments that students who regulate their cognition tend to perform well in problem-based learning (Hmelo-Silver 2004; Rozencwajg 2003), expert learning (Bransford et al. 2000; Sternberg 1998), and overall academic achievement (Peverly et al. 2003; Vrugt and Oort 2008; Winston et al. 2010).

There have been concerns, however, that these positive correlations between general metacognitive skills and academic achievement have not been as strong as would be expected (Cromley and Azevedo 2006; Jacobse and Harskamp 2012; Schunk 2008; Sperling et al. 2004; Veenman 2011). Part of this weak correlation may be because individuals who report infrequent use of metacognition place at either end of an achievement scale; that is, in addition to lower achievers who infrequently engage in metacognitive thinking, there are higher achievers who have already automatized their metacognitive scripts and therefore report less frequent use (Brown 1987; Veenman et al. 2005). Correlations may also be moderated by unrelated variables such as stereotype threat or test anxiety (Dent and Koenka 2015). Another plausible source of weak correlations may be a lack of quality instruments used to measure metacognition.

Measuring metacognition

To measure learners' metacognitive awareness and regulation, researchers have used various types of instruments, including self-report questionnaires, coded observations, think-aloud protocols, performance ratings, and interviews (Dinsmore et al. 2008; Winne and Perry 2005). In Dinsmore et al.'s (2008) review of 123 studies measuring metacognition, self-report questionnaires constituted 24% of the instruments used, exceeded in frequency only by performance ratings (31%). Self-report questionnaires are cost-effective, amenable to large-scale studies, and are typically easy to administer and score. In planning evaluations of educational interventions, for instance, evaluators are often tasked with measuring multiple proximal and distal outcomes. For practitioners with limited resources or limited access to students' class time (e.g., Young and Fry 2008), self-report questionnaires are the instrument of choice.

However, among metacognition researchers, self-report questionnaires are the most controversial class of instruments. Many scholars (Boekaerts and Corno 2005; Cromley and Azevedo 2006; Jacobse and Harskamp 2012; Tobias and Everson 2000) argue against their use for a variety of reasons pertaining to the validity of score interpretations. As with most self-report questionnaires, respondents tend to go through a process of comprehending prompts,

recalling relevant events, filling in memory gaps, and mapping their responses to the question's response scale (Tourangeau et al. 2000). With this, there is the potential for biases, such as acquiescence or social desirability, to contribute to systematic variability. In discussing metacognition inventories, Veenman (2011) points out that respondents may differ from each other in the reference points they choose and, within individuals, in their reference points across items and occasions. As with nearly any psychological instrument, metacognition questionnaires sample respondents' behaviors, which vary over occasions and contexts (Winne and Perry 2005). Although metacognition questionnaires are easy to administer, which usually results in large sample sizes that in turn reduce sampling error, when this variability is systematic across groups or time, it threatens the validity of the claims researchers and evaluators make about the relationships between metacognition and other variables or about intervention efforts.

Schellings and Van Hout-Wolters (2011) examined the pros and cons of self-report measures of metacognition instruments and concluded they are valuable for practicality and for large-scale use, and that what is needed is research that investigates the characteristics of self-report instruments for the purpose of improving them or developing new measures. One of the problems they pointed out, which is not unique to metacognition, is that correlation estimates in research using these instruments are likely explained in part by background factors. In other words, beyond the metacognition construct(s) being measured, latent nuisance variables are contributing to the variability in the metacognition scores. From a similar perspective, Berger and Karabenick (2016) stated that the field is best served by efforts to improve self-report instruments. Given the prevalence and practicality of these instruments (Berger and Karabenick 2016; Dinsmore et al. 2008), it is not likely that they will simply go away. Attention should be directed, therefore, toward research to improve these measures, particularly in examining construct-irrelevant variance.

Three frequently used self-report instruments include the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich and de Groot 1990), the Learning and Study Strategies Inventory (LASSI; Weinstein et al. 1987), and the Metacognitive Awareness Inventory (MAI; Study 1 of Schraw and Dennison 1994). The MSLQ and LASSI inventories include metacognition as subscales within the broader construct of learning strategies. The MAI was developed specifically to address the two theoretical components (or dimensions) of metacognition: knowledge of cognition (17 items) and regulation of cognition (35 items). Using established theory (Brown 1987; Flavell 1979; Jacobs and Paris 1987), Schraw and Dennison (1994) created these two subscales from a larger pool of items they developed to measure the subcomponents theorized to constitute each, with the knowledge dimension including items addressing declarative, procedural, and conditional knowledge, and the regulation dimension including items addressing planning, information management strategies, monitoring, debugging strategies, and evaluation. The MAI items with their corresponding theoretical dimensions and subcomponents are displayed in Table 9 of the Appendix.

In the original version of the MAI, the response format was a visual analogue scale from *false* to *true*, where respondents mark a slash at the location on a 100 mm line between the two end points. The subscale score is then calculated as the mean millimeter response across items. In subsequent research with the MAI, this response format has seldom been used (exceptions are Magno 2008, 2010). As displayed in Table 10 in the Appendix, studies have varied in their response-scale formats, with many using Likert-type scales. Although switching to a Likert-scale format is not problematic given that there is evidence that fully-labeled Likert-type items are as reliable as and easier to respond to than visual analogue scales (Couper et al. 2006), few

studies report using fully labeled scales (exceptions are Hughs 2015; Pucheu 2008). The use of fully labeled scales reduces satisficing among respondents, ameliorating one validity threat to the score interpretation argument (Krosnick and Presser 2010). Furthermore, the constructs of the response scales differ across studies, with some using the original false-to-true continuum and others using degrees of agreement or frequency. Although these studies provided no explanation for abandoning the original response-scale construct, a plausible explanation is that the false-to-true continuum is not appropriate for the item prompts. What does it mean for a behavior or activity to be false, true, or partially true?

Studies have also varied in how they aggregate items into composite scores. There appear to be four types of scores used in practice. Study 2 of Schraw and Dennison (1994), for example, used the best items from their small-sample ($N = 197$) exploratory factor analysis to calculate the two raw subscale scores (with 25 knowledge items and 27 regulation items), even though these items only roughly aligned with the theoretical subscales they presented in Study 1. This same scoring appeared in the second author's subsequent research (Sperling et al. 2004). Other studies have used Schraw and Dennison's theoretical subscale scoring with the two factors (Magno 2008; Young and Fry 2008) or the eight factors represented by the subcomponents (Magno 2010; Umino and Dammeyer 2016). Some (Hartley and Bendixen 2003; Stewart et al. 2007) did not specify which items they selected to aggregate into subscale scores or simply used a single-dimension score (Coutinho 2007; Kleitman and Stankov 2007; Turan et al. 2009). Still others (e.g., Hughs 2015; Pucheu 2008) have scored the MAI in three different ways, with the one-, two- and eight-subscale scoring. Whereas the theory of what constitutes metacognition has been fairly stable across studies using the MAI, the way it has been operationalized has not coalesced. What the field needs, therefore, is empirical evidence that can inform future studies' scoring procedures and provide evidence about the validity of their research claims.

Factor structure of metacognition instruments

In validity arguments about the inferences drawn from psychological instruments, one of the sources of evidence is the internal structure of the data from the responses (AERA, APA, and NCME 2014). This typically involves examinations of (a) the interrelationships among items, using item-response theory (IRT) or confirmatory factor analysis (CFA), to determine the degree to which an instrument functions in the same manner as that specified by theory, and (b) the stability of the instrument's structure across groups or time, using analyses of measurement invariance and differential item functioning (AERA, APA, and NCME 2014). With metacognition instruments similar to the MAI, this type of validity evidence has been investigated using CFA (e.g., Olejnik and Nist 1992, with the LASSI; Tock and Moxley 2017, with the MSLQ), whereas research examining IRT or measurement invariance is rarely cited or perhaps nonexistent. Exploratory factor analysis has also been used in metacognition research to examine an instrument's internal structure even though this technique is more appropriate for discovering a factor structure, such as in the instrument development phase, than for appraising the validity of its scoring inference (Brown 2006).

In CFA, the focus is on how well a measurement model, which operationalizes the theoretical factor structure, fits a set of empirical data from questionnaire responses. This is often evaluated using absolute indices such as the comparative fit index or the Tucker-Lewis index. On these, a value of 1.00 suggests that there is a perfect fit and values above .95 are commonly interpreted to mean that the model-data fit is good (Hu and Bentler 1999; cf., Yuan

et al. 2016). Competing theories of factor structure can also be evaluated using model comparisons. With the MAI, CFA can be used to compare the four measurement models that correspond with the four scoring procedures commonly used by researchers.

Whereas CFA is valuable for examining overall model fit, IRT focuses on the functioning of the items, such as how well each item fits the model and how difficult each item is to endorse. An item's fit index can reveal whether or not it functions similarly to the other items on its factor. With Likert-type questionnaire data, a polytomous IRT, such as the partial credit model (Masters 1982), can also provide information about the difficulty of endorsing one category (such as *Not very typical of me*) over another (such as *Not at all typical of me*). Unlike CFA models, standard IRT models treat the data as unidimensional, as comprising a single factor. This would be appropriate with the MAI if it were specified to measure a single general metacognition factor. An extension of the partial credit model is the multidimensional random coefficients multinomial logit (MRCML) model (Adams et al. 1997; Briggs and Wilson 2003), which can be used to analyze Likert-type response-scale data specified to be measuring more than one factor. If the MAI is specified to measure multiple factors, but does not have good CFA fit, an MRCML model can reveal items that do not fit their respective factors.

Exploratory factor analysis has been the primary tool for investigating the MAI's factor structure. In introducing the MAI, Schraw and Dennison (1994, Study 1) examined the instrument's factor structure using exploratory factor analysis with 197 Midwest U.S. college students. They found that six factors appeared to be present in the data but that these did not align with the theory that informed their eight subcomponents. They subsequently examined a constrained exploratory factor analysis model so that only two factors were permitted and found that the items moderately aligned with the knowledge and regulation dimensions, but with some discrepancies. Even though some of the items were not intended to measure the dimension they loaded on, the authors used them in calculating the raw subscales in their follow-up studies (Schraw and Dennison 1994, Study 2; Sperling et al. 2004). Their item-factor assignments are presented in Table 9.

Subsequent research on the factor structure of the MAI has yielded inconclusive results. Muis et al. (2007) made slight modifications to the regulation component and conducted a CFA but could not identify the model. Magno (2010) compared two structural equation models, one specifying the MAI as the eight theorized factors from Schraw and Dennison and the other specifying it as the two knowledge and regulation factors. He claimed that the eight-factor model fit better, but the reported Akaike and Bayesian information criteria were smaller with the two-factor model, which suggests the opposite finding; additionally, because these were structural models with many other variables, rather than measurement (CFA) models, the evidence provides little information for other researchers and practitioners seeking to know whether to calculate eight or two subscale scores when using the MAI.

Most of the remaining work investigating the factor structure of the MAI has been in translated versions, including in Turkish (Akin et al. 2007), Portuguese (Lima Filho and Bruni 2015), and Chinese (Teo and Lee 2012). Akin et al. (2007) conducted an exploratory factor analysis and claimed to have arrived at an eight factor solution matching Schraw and Dennison's (1994) intended theoretical structure; they did not conduct a CFA, however. Lima Filho and Bruni (2015) conducted CFA on the two factors, knowledge and regulation of cognition, following Schraw and Dennison's theoretical structure, but did not report the fit indices. Teo and Lee (2012) conducted an exploratory factor analysis and a follow-up CFA and argued for a three-factor model, comprising 21 of the 52 items. They did not compare their model with the original two-factor or eight-factor theoretical models, nor did they provide a

theoretical rationale for the three factors that emerged. Beyond these studies, it appears that evidence supporting or contradicting the two-dimensional and eight-dimensional models of the MAI is not discussed or does not exist in the research literature. Given the popularity of this instrument, its strong foundation in theory, and the lack of information on its factor structure, more research into its internal structure is warranted.

Research questions

The MAI was developed from a sound theoretical framework and has been extensively used. Unfortunately, there has not been adequate evidence to support decisions in how to score the responses, which has resulted in inconsistent scoring practices and questions about the validity of the various scoring inferences. Empirical evidence about its factor structure can address this need. For this reason, we conducted three studies, asking following questions:

1. Of the four scoring models frequently used with the MAI, which functions the best in explaining the pattern of responses in a set of empirical data?
2. If the best model in Study 1 does not adequately fit the empirical data, what is a subset of items, if one can be found, that constitutes the factor structure and has good fit?
3. If a well-fitting model can be found in Study 1 or 2, does it exhibit measurement invariance between groups and over time?

The four scoring models in Study 1 correspond to what has been done in practice. That is, we set out to examine a unidimensional (one factor) model, two two-dimensional models, and an eight-dimensional model. Evidence for the unidimensional model would support scoring the responses as a single general measure of metacognition; alternatively, it could indicate that there is an instrumentation effect stronger than the theorized dimensions. For the first two-dimensional model, we examine the factor structure based on Schraw and Dennison's exploratory factor analysis results; evidence for this would support their scoring procedures. For the second two-dimensional model, we examine the factor structure corresponding to the two theoretical dimensions, knowledge and regulation of cognition. For the eight-dimensional model, we examine the structure corresponding to the eight subcomponents: declarative knowledge, procedural knowledge, conditional knowledge, planning, information management strategies, monitoring, debugging strategies, and evaluation. Evidence for either of these two latter models would support scoring the MAI following the prevailing theory.

In Study 2, we follow Kline's (2011) recommendation to examine factor structure using both CFA and IRT. CFA can provide estimates of a model's overall (global) fit and can reveal error correlations among items that are unexpected. In other words, after the variability in an item is explained by its factor, there is still some error variance and when two items' errors covary, it indicates that they share some kind of relationship beyond that explained by the factor structure. Whereas some IRT models can be very similar to CFA (such as Samejima's 1969 graded response model) and yield redundant information about item fit, an MRCML model, which has foundations in the Rasch tradition (Briggs and Wilson 2003), imposes different restrictions on how items are modeled to fit their respective factors. Because researchers who use the MAI may differ in which analytic approach (CFA or Rasch) they prefer, we analyze item functioning using both CFA and MRCML modeling to identify a set of items that fit well for either type of analysis.

In Study 3, we go beyond identifying an optimal set of items and determine whether differences among groups or changes over time are attributable to the factors in the model. Evidence against measurement invariance would indicate that construct irrelevant variance is a problem; that is, it would suggest that group differences or changes over time on the MAI are at least partially due to some undefined variables that are not related to the construct.

General method

Instrument

The instrument included a question asking which gender the respondent identified with and the 52 prompts from the MAI, each with a five-point response scale, where *1* = Not at all typical of me, *2* = Not very typical of me, *3* = Somewhat typical of me, *4* = Fairly typical of me, and *5* = Very typical of me. This fully-labeled response scale differed from the original response scale of the MAI, which was a false-true semantic-differential format. This change aligns with updated research on survey scales (Krosnick and Presser 2010) in two ways. Compared to semantic-differential scales, fully labeled scales usually yield less satisficing because they are easier to respond to and provide meaning at different locations in the scale. Secondly, the typical-of-me scale more closely matches the constructs in the item prompts. Compared to scales using agreement or false-true continuums, this scale should also mitigate acquiescence bias (Krosnick and Presser 2010).

Data

We used existing data from a previous study (Vallin 2017) that met ethical standards through institutional-review-board approval; that study's purpose was to examine the effect of a strategies-training intervention in an interrupted time series design. The data in the present study comprised the MAI responses during the first two pre-intervention time points. Prior to analyses, the data were de-identified.

The response data were from 622 students in seven intact undergraduate upper-division biology and women's studies courses in a public university in Hawai'i. The class sizes ranged from 10 to 186, with a median of 61. About two-thirds ($n = 418$) of the students self-identified as female and a quarter ($n = 168$) as male; 14 identified outside of the two categories and 22 withheld a response about their gender.

In the data collection (Vallin 2017), the students had been invited to complete the questionnaire during class, early in the semester. Two modes of the instrument were administered. One was in the traditional paper-based format, administered in five classes ($n = 258$). The other was via iClicker software in two classes ($n = 364$) that had used this process in regular instruction. In this mode, the students viewed the questions on a slideshow while responding with their hand-held iClicker devices, which automatically recorded responses into a spreadsheet through a wireless network. The paper-based responses had been manually entered into a spreadsheet. With the two classes responding by iClicker mode, the instrument was administered a second time, three weeks after the first administration. Of the 364 students completing it the first time, 317 had also completed it a second time. In the present study, this second time point of data was used in the second part of Study 2 as a verification data set. Both time points of data were used in Study 3 to test measurement invariance over time.

Analyses

In the CFA analyses, we used Mplus 7 (Muthén and Muthén 2012) with full-information estimation methods, including weighted least squares means and variance adjusted (WLSMV) or maximum likelihood estimation with robust standard errors. Full-information CFA accounts for item-level missing data, which eliminates the need to drop cases with missing item responses, and permits the modeling of ordinal-level data. That is, we modeled the data as ordinal and did not eliminate cases. In the MRCML analyses, we used ConQuest 3 with the Monte Carlo estimation method (Adams et al. 2012).

Study 1

Analysis

We conducted four CFAs, one for each model for the purpose of comparing it with the other three. The first was the unidimensional model, which specified the responses on all 52 items as being explained by a single factor. The second model, following Schraw and Dennison's exploratory factor analysis results, included two factors, with 25 items explained by Factor 1 and 27 by Factor 2. The third model, based on the theoretical structure presented in Schraw and Dennison, also included two factors, with 17 items explained by a knowledge factor and 35 explained by a regulation factor. The fourth model comprised eight factors, with each specified to explain the responses on the items that constitute the eight theoretical subcomponents presented in Schraw and Dennison. The item-factor assignments for Models 2–4 are presented in Table 9.

To evaluate the fit of the models, we used criteria recommended in Hu and Bentler (1999), where adequate models typically exceed .90 on the global comparative fit index (CFI) and the Tucker-Lewis index (TLI), and well-fitting models typically have CFI and TLI estimates greater than .95, with the root mean square error of approximation (RMSEA) less than .06.

To compare the models, we used maximum likelihood estimation and compared the Akaike and Bayesian information criteria (AIC and BIC), where models with lower estimates on each are presumed to fit the data better. These indices are appropriate because the models include the same observed variables and data. Regular chi-square difference tests are not appropriate with ordinal data. Furthermore, because the second and third models included the same number of parameters, we could not use Mplus's difftest function (Muthén and Muthén 2012) with WLSMV estimation.

Results

Table 1 displays the global fit indices of the models under comparison. The CFI and TLI estimates of the first three models indicate that none met the criteria for adequate fit. The eight-dimensional model did not converge into an acceptable equation, precluding its comparison with the other models. A Haywood case was evident in its output, with the correlation between the procedural and conditional knowledge factors estimated as $r = 1.10$ and that between the monitoring and evaluation factors as $r = 1.03$. These high estimates, though they exceeded logical values, suggest that an alternative model such as the knowledge-and-regulation model will likely explain the data better.

The model comparison analysis, displayed in Table 2, indicate that the unidimensional model fit worse than the two two-dimensional models. The two-dimensional model based on Schraw and

Table 1 Fit estimates of the MAI scoring models

Model	Chi-Square (*df*)	CFI	TLI	RMSEA
Unidimensional	3634.00 (1274)	.832	.825	.055
Schraw & Dennison's EFA	3424.19 (1273)	.847	.841	.052
Knowledge & regulation	3363.28 (1273)	.851	.845	.051
Eight dimensional[a]	—	—	—	—

The estimation method was weighted least squares means and variance adjusted (WLSMV). $N = 622$; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation. [a] The eight-dimensional model did not converge

Dennison's exploratory factor analysis fit worse than the knowledge-and-regulation model. Thus, the theoretical knowledge-and-regulation model was retained. Although its two factors correlated strongly ($r = .84$), the model comparison results indicated that this theoretical two-dimensional model functioned best in explaining the pattern of responses in the empirical data.

Study 2

Analysis

We conducted Study 2 on the knowledge-and-regulation model based on the results of Study 1, which suggested it was the best model, albeit with poor fit. To identify an optimal set of items, we examined multiple types of information about item functioning using simultaneously run MRCML and CFA models in multiple iterations. Until we arrived at an optimal set, we eliminated up to three items per iteration.

In judging item functioning, we examined (a) the MRCML-estimated item fit indices; (b) the error correlations among items and between items and factors, as estimated in the CFA modification indices; and (c) the global CFA fit indices (CFI, TLI, and RMSEA) of each iteration. In examining the item fit indices, we included two estimates: The unweighted fit index, which is the traditional mean-square-error misfit estimate and which is often referred to as *outfit* in Rasch models (Smith 2004; Wilson 2005), and the weighted fit index (referred to as *infit*), which is less biased by outliers through an adjustment based on the item's variance (Smith 2004). In addition to misfit, in which a standardized fit index greater than 1.96 can indicate unexpected misfit, we examined over-fit, indicated by an index less than −1.96; overfitting items are estimated to fit the stochastic model better than expected and can indicate violations of item independence.

Table 2 Comparisons of the MAI scoring models

Model	No. of free parameters	-2LL	AIC	BIC	Decision
Unidimensional	259	78,371	78,889	80,037	Do not retain
Schraw & Dennison's EFA	260	78,149	78,669	79,822	Do not retain
Knowledge & regulation	260	78,089	78,609	79,761	Retain

The estimation method was maximum likelihood with robust standard errors. -2LL = −2*log-likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion. Lower AIC and BIC estimates indicate the model is estimated to fit better. The eight-dimensional model did not converge

Our criteria for an optimal subset of items aligned with those four types of information. Specifically, we set the stopping point of the iterative process when (a) at least 95% of the items had absolute-value standardized weighted and unweighted item fit indices less than 1.96, (b) no significant cross-factor error correlations existed among items and factors, and (c) the global fit indices (CFI and TLI) exceeded .95 with RMSEA less than .06.

In deciding which items to eliminate in the iterative process, we placed more weight on the item misfit estimates and on the modification indices that indicated error correlations between items and their unrelated factor. Simultaneously, we attempted to maintain item representation of the content and range of item difficulty. Redundant items were better candidates for elimination; we examined the item-difficulty maps (in MRCML) and pairs of items with high modification indices (in CFA). We maintained representativeness by retaining at least one item from each of the eight subcomponents theorized to constitute knowledge and regulation.

After identifying a well-functioning subset of items, we tested the model with a second set of data. This verification data set was composed of the responses from the 317 respondents during the second time point. To test the model, we examined the global fit indices using the CFA analysis and the item fit indices using the MRCML analysis.

Results

Table 3 displays the items with questionable functioning and those flagged for elimination in each iteration. The model progressively improved after several iterations of item elimination. At the 17th iteration, we met our criteria for the stopping point. In this 19-item optimal model, the global fit indices indicated good model fit (CFI = .959, TLI = .954, RMSE = .046), though the chi-square test was still significant (chi-square = 352.80, $df = 151$, $p < .001$). None of the items exhibited poor fit or correlated with items or factors in an opposing dimension. Three pairs of items had correlated errors, indicating some degree of item dependence, though each pair was within the same dimension. Table 4 displays the optimal subset of items with their dimensions (knowledge or regulation), subcomponents, category thresholds, difficulty estimates, and fit indices according to the MRCML model.

With the verification data set, composed of 317 respondents' answers during the second time point, the fit of the 19-item optimal CFA model was adequate, with RMSE = .069, CFI = .943 and TLI = .935. In the MRCML model with these data, the maximum-likelihood-estimation reliability estimates for knowledge and regulation were .80 and .84, respectively, and none of the items over-fit or misfit the model beyond our specified criteria: The item most closely approaching over-fit was Item 44 (unweighted fit = 0.88, −1.40 standardized; weighted fit = 0.89, −1.30 standardized). The one most closely approaching misfit was Item 26 (unweighted fit = 1.13, 1.50 standardized; weighted fit = 1.06, 0.80 standardized).

Study 3

Analysis

To examine measurement invariance between groups, we split the data into two groups based on whether the respondents had completed the paper-based ($n = 258$) or the iClicker ($n = 364$) mode of the questionnaire. Because we had data on gender identity,

Table 3 Iterations of item elimination based on item functioning in MRCML and CFA models

Iteration	Num. of items	MRCML		CFA				Items selected for elimination
		Misfitting items	Over-fitting items	Correlates w/ other factor	Pairs with dependence	RMSE	CFI	TLI	
1	52	25, 19, 7, 37, 42, 17, 15, 46	30, 49, 36, 33, 13, 50, 40, 23, 27, 44, 51	7, 13, 49	51&52 7&17 31&39	.051	.851	.845	7, 25, 30
2	49	19, 37, 42, 17, 31, 48, 15, 46	49, 33, 13, 36, 27, 40, 44	13, 5, 17	51&52 31&39 4&45	.052	.856	.850	13, 19
3	47	37, 42, 17, 31, 48, 15	49, 33, 36, 27, 50	5, 17, 28	51&52 31&39 4&45	.052	.860	.854	5, 17, 37
4	44	42, 31, 48, 15	49, 27, 33, 36, 35	28, 22, 49	51&52 31&39 4&45	.055	.862	.855	28, 45, 52
5	41	42, 31, 4, 48, 15	49, 33, 27, 36	41, 49, 22	31&39 22&many 2&38	.052	.884	.878	22, 31
6	39	42, 48, 4, 15	49, 33	49, 38, 23	38&many 49&many 23&many	.050	.895	.889	38, 42, 49
7	36	48, 4, 15	33, 27	34, 21, 18	15&46 11&23 2&11	.049	.906	.901	11, 15, 48
8	33	4, 47, 29, 46	33	23, 34, 18	2&23 4&21 4&39	.048	.921	.915	4, 23
9	31	47, 2, 46	33, 27	34, 3, 18	36&50 18&40 2&18	.046	.932	.927	18, 47
10	29	2, 29, 46	27, 33	34, 3, 41	3&9 36&50 14&26	.046	.935	.930	2
11	28	46, 29	33	34, 3	3&9	.047	.935	.930	3

Table 3 (continued)

Iteration	Num. of items	MRCML		CFA					Items selected for elimination
		Misfitting items	Over-fitting items	Correlates w/ other factor	Pairs with dependence	RMSE	CFI	TLI	
					14&26 36&50				
12	27	46, 29	33, 27	34, 41, 14	14&26 8&36 36&50	.047	.939	.934	14, 46
13	25	None	33	34, 41	8&36 36&50 12&21	.048	.942	.936	34, 36
14	23	29	33	None	12&21 9&10 1&8	.047	.948	.942	9, 29
15	21	12	None	None	12&21 1&8	.048	.952	.946	12
16	20	None	None	None	1&8 10&16 43&44	.048	.955	.949	1
17	19	None	None	None	6&8 10&16 43&44	.046	.959	.954	None

The items classified as misfitting had standardized misfit estimates (weighted or unweighted) greater than 1.96; over-fitting items had estimates less than −1.96; the items are reported in order of magnitude of the index. The items listed in the columns labeled *Correlates w/ other factor* and *Pairs with dependence* are the three most salient items in that iteration. In the final iteration, the MRCML (maximum-likelihood-estimation) reliability was .78 for knowledge and .82 for regulation

Table 4 MRCML difficulty estimates and fit indices of the optimal subset of items

Item	Sub-dimension	Category threshold difficulties				Item difficulty	Unweighted fit index (standardized)		Weighted fit index (standardized)	
		τ1	τ2	τ3	τ4					
Knowledge of cognition dimension										
10	DK	−3.01	−1.82	−0.14	1.93	−0.76	0.95	(−0.9)	0.95	(−0.9)
16	DK	−2.86	−1.82	−0.46	1.18	−0.98	1.09	(1.5)	1.08	(1.5)
20	DK	−3.35	−2.11	−0.46	1.05	−1.21	1.11	(1.8)	1.08	(1.4)
32	DK	−3.31	−2.09	−0.63	0.95	−1.27	0.99	(−0.1)	1.00	(0.1)
27	PK	−3.15	−1.36	0.28	1.81	−0.60	0.90	(−1.8)	0.90	(−1.9)
33	PK	−2.62	−1.57	−0.20	1.46	−0.73	0.91	(−1.7)	0.91	(−1.6)
26	CK	−2.32	−1.42	−0.44	0.75	−0.86	1.05	(1.0)	1.05	(0.9)
35	CK	−2.85	−1.44	0.35	2.31	−0.41	0.91	(−1.6)	0.91	(−1.6)
Regulation of cognition dimension										
6	P	−2.44	−0.97	0.30	1.74	−0.34	1.04	(0.6)	1.02	(0.4)
8	P	−2.20	−1.01	0.10	1.34	−0.44	1.04	(0.8)	1.04	(0.8)
39	IMS	−3.29	−1.79	−0.50	0.83	−1.19	1.06	(1.0)	1.06	(1.2)
41	IMS	−2.01	−1.19	−0.12	1.11	−0.55	1.09	(1.5)	1.08	(1.4)
43	IMS	−3.10	−1.76	−0.13	1.36	−0.91	0.96	(−0.6)	0.95	(−0.9)
21	M	−1.95	−0.54	0.58	2.23	0.08	1.04	(0.8)	1.03	(0.6)
24	E	−1.70	−0.47	0.70	2.05	0.15	0.99	(−0.2)	0.99	(−0.3)
50	E	−2.33	−0.59	0.36	1.72	−0.21	0.98	(−0.2)	0.98	(−0.4)
40	DS	−2.58	−1.61	−0.10	1.41	−0.72	0.90	(−1.8)	0.90	(−1.8)
44	DS	−3.19	−1.99	−0.56	1.45	−1.07	0.94	(−1.0)	0.93	(−1.2)
51	DS	−2.42	−1.74	−0.64	0.65	−1.03	0.95	(−0.9)	0.94	(−1.0)

The two dimensions and sub-components are based on those specified by Schraw and Dennison (1994). The threshold and item difficulty estimates are in log-odds-ratio units (logits) from the MRCML partial credit model. The standard errors were lower than 0.05. The standardized unweighted fit index is the traditional mean-square-error estimate of item misfit, and is sensitive to outliers. The weighted item fit index is adjusted to be less sensitive to outliers. Standardized indices greater than 2.0 would indicate significant misfit from the model; values less than −2.0 indicate greater than expected fit and suggest a possible violation of local independence. DK = declarative knowledge, PK = procedural knowledge, CK = conditional knowledge, P = planning, IMS = information management strategies, M = monitoring, E = evaluation, DS = debugging strategies

we also compared females (n = 418) with non-females (n = 204). Using Mplus and procedures for testing invariance with ordinal data in CFA models (Millsap and Yun-Tein 2004; Muthén and Muthén 2012), we examined the fit of the 19-item knowledge-and-regulation specification in configural, metric, and scalar models, then conducted model comparisons between the metric and configural models and between the scalar and the other two models. We used Mplus's difftest function with WLSMV estimation to compare models, setting alpha at .05 as the decision rule for rejecting the hypothesis that a more restrictive model fits as well as the less constrained one. The difftest function provides a corrected chi-square difference test to handle models with ordinal data (Muthén and Muthén 2012). A configural model specifies that for the two separate groups the same two factors explain the same items but that the factor slopes and category thresholds can differ between the two groups (where slopes are loadings and thresholds are the difficulty of endorsing one Likert-scale category over an adjacent category). It serves as the baseline for comparing metric and scalar models, which are used for assessing metric (or weak) and scalar (or strong) invariance. The metric model constrains the two groups' item slopes (or factor loadings) to be the same. The scalar model constrains the slopes and thresholds to be the same across the two groups. In

comparing this model to the two less restrictive models, we test for measurement invariance. Because perfect scalar invariance can be difficult to achieve, our contingency plan was to examine the modification indices for evidence of any threshold parameters functioning differently between the two groups, which would then lead us to examine partial scalar invariance by freeing thresholds.

For additional evidence of between-group invariance and to identify any problem items, we examined the same data in the MRCML model with tests of differential item functioning (DIF) using the Mantel Haenszel statistic procedure in ConQuest (Adams et al. 2012). This procedure reveals whether each category threshold of each item passes the ETS scores of *negligible*, *moderate*, or *severe* DIF status. If the CFA scalar invariance is not achieved, these results can be used, along with the CFA modification indices, to identify items with thresholds that are not invariant.

To test measurement invariance over time, we compared configural, metric, and scalar models using the longitudinal data of respondents who completed the MAI on the two occasions (n = 317). We applied the same model comparison procedures as that in the between-group invariance analyses, but instead of constraining parameters between groups, we constrained them between the two time points. To examine DIF in ConQuest, for the purpose of identifying problematic items, we treated the two time points as separate groups and estimated the Mantel Haenszel statistic for each threshold, as we did with the between-group DIF analyses.[1]

Results

The fit indices of the models examined for between-group measurement invariance are displayed in Table 5. The configural models indicate that when the two groups are considered, the measurement models fit well. The fit of the metric and scalar models appear to be better than their less constrained counterparts. The tests of between-group invariance, based on Mplus's difftest function, provided evidence in support of the scalar model, as displayed in Table 6. The Mantel Haenszel tests revealed no instances of DIF in either way of parsing the data into two groups. All items' thresholds were estimated as having negligible DIF.

In examining measurement invariance over time, we found that the configural, metric, and scalar models had adequate fit, as displayed in Table 7. On the CFI index, the scalar model had worse fit than the configural and metric models. The model comparison results indicated that metric invariance held, but that scalar invariance did not (Table 8). The DIF analysis revealed that Items 16 and 50 had thresholds with severe and moderate DIF, respectively. The modification indices also revealed that freeing the thresholds on these two items would improve the fit. After freeing these and running an early version of the partial scalar model, three more items (Items 33, 39, and 44) showed modification indices that supported their thresholds also being freed. After freeing all five items' thresholds, we found that the partial-scalar-invariant model, with 14 of the 19 items' thresholds restricted to be the same over time, was not significantly worse than the baseline or metric models. Two of the items with unconstrained thresholds, Items 16 and 33, became slightly easier to endorse in the second time point, whereas two, Items 39 and 44, became more difficult to endorse. For Item 50, the

[1] Although treating the two separate time points as different groups violated the assumption of local independence, the DIF part of this procedure was only for diagnosing plausible problem items rather than for making inferential decisions.

Table 5 Fit estimates of the CFA models for evaluating between-group invariance

Model	Chi-square	*df*	CFI	TLI	RMSEA
Groups: paper (n = 258) and iClicker (n = 364)					
Configural	515.64	302	.958	.952	.048
Metric	513.20	319	.962	.959	.044
Scalar	565.87	374	.962	.965	.041
Groups: female (n = 418) and not female (n = 204)					
Configural	526.73	302	.955	.949	.049
Metric	535.74	319	.956	.953	.047
Scalar	576.87	374	.959	.962	.042

The configural model specified the two groups as having equal form, but with loadings and thresholds allowed to vary. The metric model imposed the two groups have the same factor loadings. The scalar model imposed the factor loadings and item thresholds be equal

categories became more spread out; that is, the lower categories became easier to endorse and the highest category became more difficult to endorse.

Discussion

Study 1: The scoring model

Although the MAI has been widely used in research on metacognition, there has been limited research on the fidelity of its intended factor structure pertaining to how it should be scored. Similar to arguments presented in previous work (Schraw 1998; Schraw and Dennison 1994), our results support the conclusion that the 52 items function better as two theoretical dimensions, knowledge and regulation, than as a single dimension. Even though the two dimensions correlated strongly, the factor structure better explained the empirical data than did that of the unidimensional model. We also found that this theoretical structure fit better than that based on Schraw and Dennison's exploratory factor analysis, which places into question scoring procedures based on that structure.

We were unable to judge the quality of the theoretical eight-factor structure, as the model did not converge. This may have been due to poor match between the specification and the

Table 6 Tests of between-group invariance

Model comparison	Chi-square	*df*	*p*	Decision
Groups: Paper (n = 258) and iClicker (n = 364)				
Metric against configural	15.42	17	.565	Retain, test scalar
Scalar against configural	85.11	72	.138	Retain
Scalar against metric	72.36	55	.058	
Groups: Female (n = 418) and not female (n = 204)				
Metric against configural	22.27	17	.175	Retain, test scalar
Scalar against configural	86.08	72	.123	Retain
Scalar against metric	66.89	55	.131	

Difference tests were conducted using Mplus's difftest function. A p-value > .05 indicates no evidence for rejecting the more invariant model

Table 7 Fit estimates of the CFA models for evaluating time invariance

Model	Chi-square	*df*	CFI	TLI	RMSEA
Configural	1008.78	640	.936	.930	.043
Metric	1010.49	657	.939	.935	.041
Scalar	1098.49	712	.933	.934	.041
Partial scalar	1048.80	697	.939	.939	.040

The configural model specified the two time points to have the same factor structure, but with item slopes and thresholds permitted to differ in the two time points. $N = 317$. In the partial scalar model, five items' thresholds were allowed to differ between the two times: Items 16, 33, 39, 44, and 50

empirical data, or it could have been due to the complexity of the model in relation to the number of respondents. The Heywood case, with some of the sub-dimensions being very closely related, hints that it was likely the model's specification. If the results were to be trusted, it would suggest that the theoretical distinction between the procedural and conditional knowledge sub-dimensions and between the monitoring and evaluation sub-dimensions are too nuanced for this instrument. Without more model-comparison research, practitioners should be cautious about scoring the responses as eight separate sub-components. If practitioners do choose to use the 52-item MAI, our Study 1 results support the decision to derive two sub-scores based on knowledge and regulation of cognition.

Study 2: Iterative subset selection

We did, however, find problems with the 52-item knowledge-and-regulation model. The global fit indices in the CFA were below the conventional criteria for adequate fit. The first iteration of the analyses in Study 2 revealed that several items' errors correlated with the wrong factor and that at least eight items had poor fit indices. The worst fitting item was Item 25, *I ask others for help when I don't understand something*. Almost no other items contained similar wording about seeking help from others, leading us to speculate that its misfit was due to an extraneous social factor. The variance of other poorly functioning items may similarly have been due to other unspecified factors unrelated to knowledge or regulation. Sub-dimensions that add noise to the model were plausible, as some of the inter-item error correlations were between items measuring similar concepts, such as those asking about time (Items 4 and 45) or how to handle multiple options in tasks (e.g., Items

Table 8 Tests of time invariance

Model comparison	Chi-square	*df*	*p*	Decision
Metric against configural	17.15	17	.444	Retain, test scalar
Scalar against configural	126.63	72	<.001	Do not retain, identify problem thresholds and test partial scalar model
Scalar against metric	127.78	55	<.001	
Partial scalar against configural	65.95	57	.195	Retain
Partial scalar against metric	52.69	40	.086	

Difference tests were conducted using Mplus's difftest function. A p-value > .05 indicates no evidence for rejecting the more invariant model. In the partial scalar model, five items' thresholds were allowed to differ between the two times: Items 16, 33, 39, 44, and 50

2, 11, and 38). It was only at the seventh iteration, after 16 items had been eliminated, that the model functioned adequately (based on the criteria that adequate models have CFI and TLI > .90). It is worth noting that 14 of these 16 eliminated items were specified by the regulation factor, suggesting that in the full 52-item instrument, there might be more construct-irrelevant variance among the regulation items than among the knowledge items. As has been discussed in the literature (Veenman et al. 2006), some features of metacognition components are probably more successfully measured with some methods than others. For research seeking to make strong claims about respondents' regulation of cognition, it would be wise to use multiple methods of measurement (Greene and Azevedo 2010; Schellings and Van Hout-Wolters 2011). If our Study 2 findings represent what occurs in the larger population, practitioners should heed caution when aggregating the responses from the 52 items into scores. The knowledge and regulation scores will likely be biased by other sources of variability among the items, which threatens the validity of the scoring inferences and therefore the validity in using the data to make claims about the effects of metacognition interventions. With the items having high error correlations, researchers might consider using parcels or testlets in computing scores. Because the MAI is long—longer than its near counterparts on subsections of the MSLQ and LASSI—item eliminations, like the ones we made, are also an option.

Our shortened version of the MAI functioned well with our data. The final iteration resulted in 19 items that had good fit in both CFA and MRCML models. With the verification data set, the CFA model fit was lower, but still classifiable as adequate. This lower fit was not surprising because the verification data set was smaller, with 317 cases. The MRCML model with the verification data revealed no misfitting items. These results suggested that the 19-item model suited both a partial-credit-model IRT scoring framework and a CFA (or graded-response-model) scoring framework. The former is valuable for researchers in the Rasch-analysis tradition. The latter is valuable for researchers seeking to include item response data in a larger structural equation model.

Our provisional conclusion is that this 19-item subset will function equally well in similar samples of university students. There is still room for further instrument improvement, however, particularly that focusing on content representation. We were careful to maintain representation of the eight sub-components, but this representation is not ideally balanced. In the knowledge of cognition dimension, declarative knowledge, which has four items, is better represented than procedural and conditional knowledge. In the regulation of cognition dimension, monitoring, planning, and evaluation have fewer items than debugging and information management strategies. Nonetheless, given the lack of credible empirical data for the eight-factor model and the trend that metacognition instruments tend to be more blunt than the fine-grained theoretical descriptions (Pintrich et al. 2000), this shortened MAI is likely appropriate for practical use.

Study 3: Invariance

The results of Study 3 suggested that the 19-item instrument is invariant across groups. The items appear to function in the same manner regardless of the mode of the instrument, whether it was by iClicker or paper, and regardless of the gender identity of the respondents. The differential-item functioning results were consistent, revealing no evidence of any item-level bias in the two between-group analyses. Other research is

required, however, before making strong claims about invariance. In our Study 2 endeavor to detect poorly functioning items, we sought as large a sample size as we could get. With limited resources in Study 3, we were unable to include an outside group, limiting the generalizability of these between-group invariance findings. Our conclusion, therefore, is that our empirical evidence provides good, albeit provisional, support for comparing groups with this shortened MAI.

Strong invariance across time was not achieved. Only partial scalar invariance was attained, with five items' thresholds unconstrained across time. This suggests that studies using this subset to measure longitudinal change may encounter variability in scores that is partially amplified or suppressed by these differentially functioning items.

The two items with thresholds that became easier to endorse in the second time point were Item 16, *I know what the teacher expects me to learn*, and Item 33, *I find myself using helpful learning strategies automatically*. The increased easiness in Item 16 makes sense, given that the second time point was several weeks into the classes' semesters and that as students become more accustomed to a class, they begin to become more aware of the teacher's expectations. As for Item 33, a logical explanation for the increase is unclear.

The two items with thresholds that became harder to endorse, Item 39, *I try to translate new information into my own words*, and 44, *I re-evaluate my assumptions when I get confused*, were both about strategies. A plausible explanation is that as students move forward in the particular class, they find themselves not needing these specific strategies, thereby not endorsing the higher ends of the typical-of-me scale.

As for the item that expanded in variability in its thresholds, Item 50, *I ask myself if I learned as much as I could have once I finish a task*, a plausible speculation is that compared to when they read the question at the start of the semester, students at the second time point may have considered recent events and tasks and responded in light of these concrete instances instead of responding with a judgment of their general behavior such as a typical instance or an average across instances. When students complete the MAI in the first time point, early in a course, they will have recently come back from break, perhaps with some time having elapsed since they engaged in academic tasks.

These speculative explanations about differential item functioning are consistent with arguments that self-reports about strategy use are dependent on specific learning events (Greene and Azevedo 2010; Samuelstuen and Bråten 2007; Schellings and Van Hout-Wolters 2011; Winne and Perry 2005). It is often pointed out that respondents' answers are typically in light of recent and salient experiences (based on Tourangeau et al. 2000). Some scholars (Samuelstuen and Bråten 2007) go so far as to advise against the practice of measuring global metacognitive strategy use altogether. At the same time, it is also maintained that these biases are present among all types of measures of metacognition, not just self-reports (Dinsmore et al. 2008; Schellings and Van Hout-Wolters 2011), and that measuring global metacognition strategy use is valuable (Schraw 1998). Researchers seeking information on metacognition as a proclivity (Schraw 2010) or disposition (Halpern 1998; Kuhn and Dean 2004) should consider the grain size of their research purpose when selecting an instrument (Berger and Karabenick 2016; Pintrich 2004; Veenman et al. 2006). With the MAI, it may be that some of the poorly fitting items or the items responsible for weakening the invariance over time are of a grain size that is more appropriate for measuring metacognition in specific events than as a propensity or domain-general set of skills.

Limitations and future directions

This study is not without limitations. The data were from a sample that was not a random draw and the participants were nested in classes in a single institution. The course topics, however, were not in a field, such as educational psychology, that would draw students with previous proclivities to studying metacognition. So, although our sample cannot constitute an unbiased estimate of the larger post-secondary population, with regard to the construct itself, this may not be a strong threat to validity. Given that the MAI is self-report, there are likely biases such as acquiescence and social desirability in effect, and medium- or high-stakes decisions should be accompanied by other measures. It is promising, though, that our results suggest that if bias exists, it is consistent between groups, which means that between-group inferential statistical tests are not likely influenced by such bias. Further research is needed to confirm the theoretical structure with our subset of items. Notwithstanding, we have examined the factor structure and invariance of items on the MAI in light of prevailing theory and have contributed to the ongoing effort to measure this elusive construct. Our findings provide evidence that the 52-item MAI is suspect, but that a subset of the items is likely useful, meriting further examination in other contexts.

The norm in self-report measures of metacognition is to solicit responses on normative response scales, such as Likert-type response scales or anchor-labeled semantic-differential scales. An alternative is to use absolute response scales, such as in forced-choice format questions. These have been discussed in the metacognition research (Pintrich et al. 2000; Winne and Perry 2005) but with cautions against their use due to the problems in modeling the ipsative responses. Forced-choice formats are typically less prone to the biases inherent in response-scale format questions, and recent work in analyzing ipsative and quasi-ipsative models (Brown and Maydeu-Olivares 2013) may lead to promising new developments in metacognition measurements. The MAI may not be an ideal measure, but the knowledge we gain from examining its factor structure can inform the ongoing quest for a feasible self-report measure of metacognition.

Conclusion

Researchers will likely continue to include self-report measures of metacognition in their research, particularly if metacognition is theorized as a component of a larger construct, such as critical thinking, or if it is included in theory-of-change models as a mediating or outcome variable. In making claims about the validity of their findings, researchers need to consider the quality of the instruments they use and the degree to which they align with theory. The MAI has been extensively used, but methods of scoring the responses have been inconsistent and have lacked adequate empirical support in light of theory. The present study informs practitioners of ways they can score the MAI and contributes to the ongoing effort to improve the measurement of this broad construct. We should not stop here. Among the features of self-report instruments targeting metacognition as a general proclivity, grain size, item dependencies, and question format deserve more research attention.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Table 9 The MAI prompts and the factor assignments for the Study 1 model comparisons

Item	Prompt	Model 2	Model 3	Model 4
1	I ask myself periodically if I am meeting my goals.	2	R	M
2	I consider several alternatives to a problem before I answer.	2	R	M
3	I try to use strategies that have worked in the past.	1	K	PK
4	I pace myself while learning in order to have enough time.	2	R	P
5	I understand my intellectual strengths and weaknesses.	1	K	DK
6	I think about what I really need to learn before I begin a task.	2	R	P
7	I know how well I did once I finish a test.	1	R	E
8	I set specific goals before I begin a task.	2	R	P
9	I slow down when I encounter important information.	1	R	IMS
10	I know what kind of information is most important to learn.	1	K	DK
11	I ask myself if I have considered all options when solving a problem.	2	R	M
12	I am good at organizing information.	1	K	DK
13	I consciously focus my attention on important information.	1	R	IMS
14	I have a specific purpose for each strategy I use.	2	K	PK
15	I learn best when I know something about the topic.	1	K	CK
16	I know what the teacher expects me to learn.	1	K	DK
17	I am good at remembering information.	1	K	DK
18	I use different learning strategies depending on the situation.	1	K	CK
19	I ask myself if there was an easier way to do things after I finish a task.	2	R	E
20	I have control over how well I learn.	1	K	DK
21	I periodically review to help me understand important relationships.	2	R	M
22	I ask myself questions about the material before I begin.	2	R	P
23	I think of several ways to solve a problem and choose the best one.	2	R	P
24	I summarize what I've learned after I finish.	2	R	E
25	I ask others for help when I don't understand something.	1	R	DS
26	I can motivate myself to learn when I need to.	1	K	CK
27	I am aware of what strategies I use when I study.	2	K	PK
28	I find myself analyzing the usefulness of strategies while I study.	2	R	M
29	I use my intellectual strengths to compensate for my weaknesses.	1	K	CK
30	I focus on the meaning and significance of new information.	1	R	IMS
31	I create my own examples to make information more meaningful.	1	R	IMS
32	I am a good judge of how well I understand something.	1	K	DK
33	I find myself using helpful learning strategies automatically.	1	K	PK
34	I find myself pausing regularly to check my comprehension.	2	R	M
35	I know when each strategy I use will be most effective.	2	K	CK
36	I ask myself how well I accomplish my goals once I'm finished.	2	R	E
37	I draw pictures or diagrams to help me understand while learning.	2	R	IMS
38	I ask myself if I have considered all options after I solve a problem.	2	R	E
39	I try to translate new information into my own words.	1	R	IMS
40	I change strategies when I fail to understand.	2	R	DS
41	I use the organizational structure of the text to help me learn.	2	R	IMS
42	I read instructions carefully before I begin a task.	1	R	P
43	I ask myself if what I'm reading is related to what I already know.	2	R	IMS
44	I re-evaluate my assumptions when I get confused.	2	R	DS
45	I organize my time to best accomplish my goals.	1	R	P
46	I learn more when I am interested in the topic.	1	K	DK
47	I try to break studying down into smaller steps.	2	R	IMS
48	I focus on overall meaning rather than specifics.	2	R	IMS
49	I ask myself questions about how well I am doing while learning something new.	2	R	M
50	I ask myself if I learned as much as I could have once I finish a task.	2	R	E
51	I stop and go back over new information that is not clear.	1	R	DS
52	I stop and reread when I get confused.	1	R	DS

The item numbers, prompts, and specified factors are from Schraw and Dennison (1994). Model 1 specified all items on a single factor, and is not displayed; Model 2's factors are based on the Schraw & Dennison's two-factor exploratory factor analysis results; Model 3's factors are based on the two-dimensional theoretical model, where K = knowledge and R = regulation; Model 4's factors are based on the eight-dimensional theoretical model, where DK = declarative knowledge, PK = procedural knowledge, CK = conditional knowledge, P = planning, IMS = information management strategies, M = monitoring, DS = debugging strategies, and E = evaluation

Table 10 Studies' scoring and response-scale formatting of the MAI

Study	Scoring	Response-scale formatting		
		Type	Construct	Labeling
Magno 2010	Eight latent factors	0-to-100 mm visual analogue	*False* to *True*	End points
Akin et al. 2007	Eight subscales	5-point scale	*Always false* to *Always true*	—
Umino and Dammeyer 2016	Eight subscales	5-point scale	*Strongly disagree* to *Strongly agree*	—
Hughs 2015	Eight, two, and single subscales	5-point scale	*Always false* to *Always true*	All points
Pucheu 2008	Eight, two, and single subscales	5-point scale	*Always false* to *Always true*	All points
Schraw and Dennison 1994, Study 2	Two subscales based on EFA in Study 1	0-to-100 mm visual analogue	*False* to *True*	End points
Magno 2008	Two subscales	0-to-100 mm visual analogue	*Always false* to *Always true*	End points
Young and Fry 2008	Two subscales	5-point scale	*False* to *True*	End points
Sperling et al. 2004	Two subscales based on EFA in Schraw & Dennison, Study 1	5-point scale	—	—
Hartley and Bendixen 2003	Two subscales	5-point scale	—	—
Stewart et al. 2007	Two subscales	Scale, unspecified number of points	—	—
Lima Filho and Bruni 2015	Two latent factors	Scale, unspecified number of points	—	—
Kleitman and Stankov 2007	A single latent factor	6-point scale	*Never* to *Always*	—
RincónGallardo 2009	A single score	5-point scale	*Not true* to *True*	End points
Turan et al. 2009	A single score	5-point scale	—	—
Coutinho 2007	A single score	7-point scale	*Strongly disagree* to *Strongly agree*	—
Muis et al. 2007	Custom	0 to 100 number entry	*False* to *True*	End points
Teo and Lee 2012	Custom	7-point scale	*Strongly disagree* to *Strongly agree*	—

The dash indicates that the information was not reported in the study. EFA = exploratory factor analysis

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. doi:10.1177/0146621697211001.

Adams, R. J., Wu, M. L., Haldane, S. A., & Xun, S. X. (2012). *ConQuest (version 3.0.1) [computer software]*. Camberwell: Australian Council for Educational Research.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Akin, A., Abaci, R., & Çetin, B. (2007). The validity and reliability of the Turkish version of the metacognitive awareness inventory. *Educational Sciences: Theory & Practice, 7*, 671–678.

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching* (Vol. 32). San Francisco: John Wiley & Sons.

Berger, J.-L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment, 21*, 19–33. doi:10.1080/10627197.2015.1127751.

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

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience and school* (expanded ed.). Washington, D.C.: National Academy Press.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87–100.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*, 36–52. doi:10.1037/a0030641.

Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Weinert & R. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65–116). Mahwah: Erlbaum.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

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

Coutinho, S. A. (2007). The relationship between goals, metacognition, and academic success. *Educate~: The Journal of Doctoral Research in Education, 7*(1), 39–47.

Cromley, J. G., & Azevedo, R. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning, 1*, 229–247. doi:10.1007/s11409-006-9002-5.

Dent, A. L., & Koenka, A. C. (2015). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, Advance online publication*, *1–50*. doi:10.1007/s10648-015-9320-8.

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.

Duncan, T. G., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist, 40*, 117–128. doi:10.1207/s15326985ep4002_6.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–235). Hillsdale: Lawrence Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*, 906–911. doi:10.1037/0003-066X.34.10.906.

Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist, 45*, 203–209. doi:10.1080/00461520.2010.515935.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449–455. doi:10.1037/0003-066X.53.4.449.

Hartley, K., & Bendixen, L. D. (2003). The use of comprehension aids in a hypermedia environment: Investigating the impact of metacognitive awareness and epistemological beliefs. *Journal of Educational Multimedia and Hypermedia, 12*, 275–289.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*, 235–266. doi:10.1023/B:EDPR.0000034022.16470.f3.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. doi:10.1080/10705519909540118.

Hughs, J. A. (2015). *Impact of online self-regulated professional development on technology and engineering educators metacognitive awareness*. (doctoral dissertation). ProQuest Dissertations & Theses Global database. (accession no. 3710627).

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22*, 255–278.

Jacobse, A. E., & Harskamp, E. G. (2012). Towards efficient measurement of metacognition in mathematical problem solving. *Metacognition and Learning, 7*, 133–149. doi:10.1007/s11409-012-9088-x.

Khosa, D. K., & Volet, S. E. (2014). Productive group engagement in cognitive activity and metacognitive regulation during collaborative learning: Can it explain differences in students' conceptual understanding? *Metacognition and Learning, 9*, 287–307. doi:10.1007/s11409-014-9117-z.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161–173. doi:10.1016/j.lindif.2007.03.004.

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

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Second ed., pp. 263–313). Bingley: Emerald.

Ku, K. Y. L., & Ho, I. T. (2010). Metacognitive strategies that enhance critical thinking. *Metacognition and Learning, 5*, 251–267. doi:10.1007/s11409-010-9060-6.

Kuhn, D., & Dean Jr., D. (2004). Metacognition: A bridge between cognitive psychology and educational practice. *Theory Into Practice, 43*, 268–273. doi:10.1207/s15430421tip4304_4.

Lima Filho, R. N., & Bruni, A. L. (2015). Metacognitive awareness inventory: Translation and validation from a confirmatory analysis. *Psicologia: Ciência e Profissão, 35*, 1275–1293. doi:10.1590/1982-3703002292013.

Magno, C. (2008). Reading strategy, amount of writing, metacognition, metamemory, and apprehension as predictors of English written proficiency. *Asian EFL Journal, 29*, 15–48.

Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and Learning, 5*, 137–156. doi:10.1007/s11409-010-9054-4.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-category measures. *Multivariate Behavioral Research, 39*, 479–515.

Muis, K. R., Winne, P. H., & Jamieson-Noel, D. (2007). Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *British Journal of Educational Psychology, 77*, 177–195. doi:10.1348/000709905X90876.

Muthén, L. K., & Muthén, B. O. (2012). *MPlus user's guide* (Seventh ed.). Los Angeles: Muthén & Muthén.

Olejnik, S., & Nist, S. L. (1992). Identifying latent variables measured by the learning and study strategies inventory (LASSI). *The Journal of Experimental Education, 60*, 151–159.

Peverly, S. T., Brobst, K. E., Graham, M., & Shaw, R. (2003). College adults are not good at self-regulation: A study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology, 95*, 335–346. doi:10.1037/0022-0663.95.2.335.

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

Pintrich, P. R., & de Groot, E. V. (1990). Motivated strategies for learning questionnaire. *Retrieved from PsycTESTS.* doi:10.1037/t09161-000.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43–97). Lincoln: Buros Institute.

Pucheu, P. M. (2008). *An investigation of the relationships between the scoring rubrics inventory and the metacognitive awareness inventory as reported by secondary school core -subject teachers.* (doctoral dissertation). ProQuest Dissertations & Theses Global database. (accession no. 3313868).

RincónGallardo, T. J. (2009). *The effect of the use of learning journals on the development of metacognition in undergraduate students.* (doctoral dissertation). ProQuest Dissertations & Theses Global database. (accession no. 3389888).

Rozencwajg, P. (2003). Metacognitive factors in scientific problem-solving strategies. *European Journal of Psychology of Education, 18*, 281–294. doi:10.1007/BF03173249.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No.* 17.

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

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.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* New York: Basic Books.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*, 113–125. doi:10.1023/a:1003044231033.

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist, 45*, 258–266. doi:10.1080/00461520.2010.515936.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475. doi:10.1006/ceps.1994.1033.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*, 111–139. doi:10.1007/s11165-005-3917-8.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review, 20*, 463–467. doi:10.1007/s10648-008-9086-3.

Smith, E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 93–122). Maple Grove: JAM Press.

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *Educational Research & Evaluation, 10*, 117–139.

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science, 26*, 127–140. doi:10.1023/a:1003096215103.

Stewart, P. W., Cooper, S. S., & Moulding, L. R. (2007). Metacognitive development in professional educators. *The Researcher, 21*(1), 32–40.

Teo, T., & Lee, C. B. (2012). Assessing the factorial validity of the metacognitive awareness inventory (MAI) in an Asian country: A confirmatory factor analysis. *The International Journal of Educational and Psychological Assessment, 10*(2), 92–103.

Tobias, S., & Everson, H. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 147–222). Lincoln: Buros Institute.

Tock, J. L., & Moxley, J. H. (2017). A comprehensive reanalysis of the metacognitive self-regulation scale from the MSLQ. *Metacognition and Learning, 12*, 79–111. doi:10.1007/s11409-016-9161-y.

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

Turan, S., Demirel, Ö., & Sayek, İ. (2009). Metacognitive awareness and self-regulated learning skills of medical students in different medical curricula. *Medical Teacher, 31*(10), 477–483. doi:10.3109/01421590903193521.

Umino, A., & Dammeyer, J. (2016). Effects of a non-instructional prosocial intervention program on children's metacognition skills and quality of life. *International Journal of Educational Research, 78*, 24–31. doi:10.1016/j.ijer.2016.05.004.

Vallin, L. M. (2017). *Metacognition as a pedagogical approach to improve students' use of metacognitive strategies*. Honolulu: (Unpublished doctoral dissertation), University of Hawai'i at Mānoa.

Vauras, M., Iiskala, T., Kajamies, A., Kinnunen, R., & Lehtinen, E. (2003). Shared-regulation and motivation of collaborating peers: A case analysis. *Psychologia, 46*, 19–37.

Veenman, M. V. J. (2011). Learning to self-monitor and self-regulate. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 197–218). New York: Taylor & Francis Group.

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science, 33*, 193–211. doi:10.1007/s11251-004-2274-8.

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

Vrugt, A., & Oort, F. J. (2008). Metacognition, achievement goals, study strategies and academic achievement: Pathways to achievement. *Metacognition and Learning, 3*, 123–146. doi:10.1007/s11409-008-9022-4.

Weinstein, C.E., Schulte, A., & Palmer, D.R. (1987). *The learning and study strategies inventory* Clearwater: H&H Publishing.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Psychology Press.

Winne, P. H., & Perry, N. E. (2005). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Burlingtton: Elsevier Academic Press.

Winston, K. A., Van der Vleuten, C. P. M., & Scherpbier, A. J. J. A. (2010). An investigation into the design and effectiveness of a mandatory cognitive skills programme for at-risk medical students. *Medical Teacher, 32*, 236–243. doi:10.3109/01421590903197035.

Young, A., & Fry, J. D. (2008). Metacognitive awareness and academic achievement in college students. *Journal of the Scholarship of Teaching and Learning, 8*(2), 1–10.

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 319–330. doi:10.1080/10705511.2015.1065414.