The Psychometric Properties of the Metacognitive Awareness Inventory in the Japanese EFL Context

XETHAKIS, Larry

Abstract

The measurement properties of the Metacognitive Awareness Inventory (MAI; Schraw & Dennison, 1994) as well as three versions of the instrument prominent in the literature are reported in this study. Responses from university students at four universities in Western Japan (N = 729) comprised the data-set. Normality of score distributions for test items was examined, and reliability estimates (Cronbach's alpha) were calculated for the various subscales reported for each of the versions tested. Confirmatory factor analysis (CFA) was employed to examine the structural validity of four versions of the MAI: 1) the correlated, two-factor structure originally hypothesized by Schraw and Dennison (1994); 2) the hierarchical, three-factor structure proposed by Abe and Ida (2010); 3) the correlated, three-factor structure proposed by Niwa and Yamaji (2017); and, 4) the correlated, three-factor structure proposed by Teo and Lee (2012). All of the models performed less than satisfactorily, displaying a very poor degree of fit with the structure of the scores in the dataset. These results, by providing negative evidence for a two- or three-construct structure underlying scores on the MAI, positively assist with future theorization of the structure of metacognitive awareness.

Introduction

Metacognition, according to one influential and succinct definition, is the "ability to reflect on, understand and control one's learning," (Schraw & Dennison, 1994, p. 460). With this conception in mind, metacognitive awareness can be construed as the extent to which an individual is aware of, and can reflect upon, his/her own ideas about learning, the ways he/she goes about learning and studying, and how he/she manages these processes. Learners' metacognitive capabilities and their awareness of these capabilities have been shown to impact a wide range of subject areas and skills, including mathematics (e.g., Artz & Armour-Thomas, 1992), reading (e.g., Brown & Palincsar, 1989), language acquisition (e.g., Haukas, 2018) and collaboration with others (e.g., Cantwell & Andrews, 2002; Molenaar, Sleegers & von Boxtel, 2014).

In the realm of language acquisition more specifically, research over the past two decades has shown the important role played by metacognition and metacognitive awareness. Victori (1999), in an early study, found that good language learners employ strategies more often than poor ones. The use of strategies has been considered to be of such importance that Anderson (2002) asserted, "understanding and controlling cognitive processes may be one of the most essential skills that classroom teachers can help second language learners develop" (p. 2). In addition to its impact on the use of strategies, metacognition is central in the planning and monitoring of language learning. The ability to reflect on and manage one's learning has been shown to be important for successfully learning languages (Haukas, 2018). Metacognitive capability and awareness allow learners to reflect on their learning, and thus make better decisions on what they need to do to improve their learning (Anderson, 2008). In fact, metacognition can serve as a strong predictor of second language performance, as Raoofi, Chan, Mukundan and Rashid (2014) found in their review of studies examining metacognition and language learning.

Learners' metacognitive capabilities are also crucial for effective collaboration, which often underpins communicative activities in the language learning classroom. Students who exhibit greater awareness of their own metacognitive propensities tend to be more positively disposed to working in groups (Cantwell & Andrews, 2002). Moreover, in group-work environments, students require skills to convey their subject knowledge to other group members, while simultaneously understanding the knowledge of their peers (Smith & Mancy, 2018). Artz & Armour-Thomas (1992) found that learners exhibiting a higher level of metacognition were more successful in collaborative problem solving. This parallels a later finding by Lumma-Sellenthin (2012) that medical students with a better grasp of metacognitive strategies were better able to learn communication skills.

Metacognition is also an important factor in the organizational aspects of group work. When working in groups, learners need to negotiate roles, manage conflict, plan and monitor the accomplishment of tasks, as well as reflect upon and revise these processes as necessary (Splichal, Oshima & Oshima, 2018). In addition to the ability to organize and manage group processes, metacognition plays a role in student interaction and engagement with group members. Webb (2009) notes that "the extent to which students benefit from working with other students depends on the nature of students' participation in group work" (p. 2). Learners who are able to participate constructively in discussions, provide feedback, develop the ideas of other group members and help co-construct knowledge enhance the group's collective learning (Molenaar, et al., 2014).

Collaboration between learners is becoming a topic of increasing importance in the Japanese educational context. MEXT has placed greater emphasis on group-based learning in the language classroom over the past decade (MEXT 2009a; 2009b; 2017a; 2017b). This emphasis on group work goes beyond the language classroom and is also aimed at reforming the way all subjects are taught, by introducing group-based learning approaches into elementary, secondary and university settings (MEXT 2013, 2014a, 2014b).

The Metacognitive Awareness Inventory (MAI; Schraw & Dennison, 1994) is one of the most widely used self-report measures of metacognition (e.g., Kondo, et al., 2012; Lee, Teo & Bergin, 2009; Raes, Schellens, De Wever & Vanderhoven, 2012; Raoofi, et al., 2014; Sperling, Howard, Miller

& Murphy, 2002; Zhang, 2010). The MAI was originally conceived by Schraw and Dennison as a 52-item measure of two theorized aspects of metacognition, namely, knowledge of cognition and regulation of cognition. The authors initially hypothesized that these two categories comprised eight distinct subcomponents. Three of these—declarative, procedural and conditional knowledge—were posited as subcomponents of knowledge of cognition. The remaining five—planning, monitoring, evaluating, information management and debugging strategies—were theorized to be subcomponents of regulation of cognition.

Schraw and Dennison employed exploratory factor analysis (EFA) on a set of scores from 197 undergraduates to investigate the hypothesized structure of the instrument. Their initial EFA resulted in a six-factor solution, rather than an eight-factor solution as hypothesized. Moreover, the loading of the items on these six factors did not correspond well with the hypothesized subcomponents either. For this reason, they concluded that the structure underlying the MAI did not actually reflect the eight hypothesized subcomponents of metacognition, although from a critical point of view CFA would be a preferable method to support such an inference. Subsequent to this, they imposed a constrained two-factor solution onto the data. This EFA resulted in a more interpretable solution, with the items from the knowledge of cognition category loading primarily on the first factor, and the items from the regulation of cognition category loading primarily on the second factor. There were some inconsistencies in the loading of several items, however. Eleven of the items hypothesized to fall into the regulation of cognition category, loaded on the first, knowledge of cognition factor, and one item from the knowledge of cognition category loaded on the second factor, which was primarily concerned with the regulation of cognition. Finally, six of the original 52 items cross-loaded on both factors, while two items failed to load on either factor.

On the basis of these results, the authors concluded that the MAI measured two aspects of metacognition, which broadly paralleled the two categories described above. The two factors were termed the Knowledge of Cognition (KC; 25 items) and Regulation of Cognition (RC; 19 items) subscales. These two subscales were found to be moderately correlated ($r = .54$), and to each have a reliability estimate (Cronbach's alpha) of .91. It should be noted that in their original paper, Schraw and Dennison (1994) performed separate EFAs on two different sets of data. The content of each of the factors in the respective two-factor solutions from both of the EFAs was quite similar, however a small number of items in the second EFA failed to load on the same factor as in the first EFA. For the purposes of this paper, the factor loadings of the items in the first EFA performed by Schraw and Dennison, as described above, are employed in specifying models for the analyses outlined below.

Despite the fact that the MAI is a popular and widely used self-report measure of metacognition and metacognitive awareness, some questions remain as to the dimensionality of the instrument (Harrison & Valin, 2017). As discussed above, the MAI was originally hypothesized to measure eight dimensions of metacognitive awareness. However, the EFAs conducted by Schraw and Dennison (1994) resulted in a two-factor structure. In addition, a large number of items that were included on the KC subscale in study were originally hypothesized to belong in the regulation category, while several that loaded onto the RC subscale were supposed to measure knowledge of cognition. Further research on the factor structure of the MAI has not resolved either of these issues clearly.

Muis, Winne, and Jamieson-Noel (2007) performed CFA on an eight-factor model that displayed what could be considered an adequate degree of fit (applying the cut-off criteria of Hu & Bentler, 1999). For this model, the reported goodness-of-fit indices were the Goodness-of-fit Index (GFI) = .94, CFI = .92, RMSEA = .06 and Bentler-Bonett normed fit index (NFI) = .87. However, it must be noted that these values were for an adjusted model with five items removed (two from the planning subcomponent and three from the debugging component) due to near-zero or negative loadings, and thus Muis, et al., state that their results may not be generalizable to other populations. Magno (2010) carried out CFAs on both a two-factor and eight-factor structure, and found support for the eight-factor structure. The goodness-of-fit indices reported in this case were: CFI = .941 and RMSEA = .05. These results were obtained in a SEM, and it should be noted that this modeling strategy is aimed at understanding the relationships that hold between constructs, rather than only between indicators and constructs (i.e. the measurement model with which CFA is concerned). Thus, this result has little bearing on the measurement discussion. For this reason, the evidence for the underlying dimensionality of the MAI provided by Magno's results requires qualification, and is arguably less helpful.

A third study which has examined this question is that of Harrison and Valin (2017). They tested both the two- and eight-factor models using CFA. They found that the eight-factor model generated a Heywood case with correlations between several of the factors exceeding 1.00. Schraw and Dennison's (1994) originally hypothesized model was also tested in this study and found to exhibit poor fit. The reported values for this two-factor model were: TLI = .841, CFI = .847, and RMSEA = .052. As a result of the poor fit for this model, the authors used item-response theory to develop a 19-item model with two underlying constructs which displayed good fit (TLI = .954; CFI = .959; RMSEA = .046).

In addition to the above studies in English speaking populations, there have been a number of studies which examined the factor structure of translated versions of the MAI. Akin, Abaci and Çetin (2007) conducted an EFA on a Turkish version of the MAI and found support for eight underlying factors. Zhang (2010) found that the 52 items of the MAI loaded onto two factors in the EFA carried out in her study. A third effort to uncover the underlying structure of the MIA was that of Teo and Lee (2012), who attempted to find evidence for the originally postulated eight-factor structure of the MAI, but, in a manner similar to Schraw and Dennison, were unsuccessful in producing an interpretable solution. Further EFAs revealed a three-factor structure comprising 45 items. When this model was tested with CFA, it was found to fit the data poorly (TLI = .756; CFI = .768; RMSEA = .076; SRMR = .068). After modification, the number of items was reduced to 21 and this model displayed a just adequate degree of fit (TLI = .903; CFI = .914; RMSEA = .063;

SRMR = .048). It must be noted here that the authors did not clearly state whether the results of the CFAs were based on a new set of data, or whether they were based on the same data used in the EFA. If the latter, it could be expected that the CFA would re-produce an acceptable degree of fit, because the confirmatory model is empirically derived from the same dataset in which it is being tested, which of course would not qualify as an a priori test. From the above, we can see that there is a greater weight of evidence suggesting that the MAI possesses two underlying dimensions. However, the possibility of eight, or even three dimensions has not been definitively ruled out.

In fact, the most commonly used (e.g., Aono & Nishino, 2016; Kondo, et al., 2012; Saito, 2016) Japanese version of the MAI (Abe & Ida, 2010) is hypothesized to possess three underlying dimensions. This version was developed employing EFA on a dataset of scores from 246 university students. The results from this initial EFA suggested a three-factor solution comprising a total of 37 items. As in Teo & Leo (2012) above, this solution was then tested using CFA. The results showed that this model did not fit the structure of the scores in the dataset well (GFI = .812; AGFI = .789; RMSEA = .053). The authors then removed 9 items which had loaded on their respective factors at less than .55 and subjected this 28-item revised model to CFA. The results of this CFA showed that the revised model displayed better fit than the initial 37-item model (GFI = .865; AGFI = .842; RMSEA = .048), but were still unsatisfactory. These three resultant factors were entitled Monitoring (11 items), Control (9 items) and Metacognitive Knowledge (8 items). The reliability estimates of these three subscales were reported as .878, .788 and .749, respectively. A moderate degree of correlation between the three factors was found by the authors, and this led them to assume that the three factors were influenced by the higher-order construct of metacognition. To the best of this author's knowledge, Abe and Ida's model for the underlying structure of the MAI is the only one which postulates a second-order construct.

Two further points should be kept in mind in any evaluation of Abe and Ida's proposed model. First, no new data was employed in the CFA to test the postulated model, and thus, as mentioned above in relation to Teo and Lee's (2012) model, there should be a reasonable degree of improvement of fit exhibited, because the model is being tested in the same set of data in which it was empirically derived. Second, and as stated above, while the 28-item model did exhibit improved fit, by no means should these values be interpreted as indicating good fit. To make a claim of good fit when using the GFI or AGFI, the model should return a value of at least .90 and preferably higher (Hair, Black, Babin & Anderson, 2014).

Abe and Ida's 28-item version was further examined by Niwa & Yamaji (2017), who employed EFA on a set of scores from 226 first-year university students. This study also found a three-factor structure underlying scores on the MAI. However, a large number of the items loaded on different factors than they had in Abe and Ida's study, while three items failed to load above a value of .35 on any factor. In order to clarify the content of the factors, the five items that loaded most strongly on each of the three factors were retained and a further EFA was undertaken. This resulted in a fifteen-item instrument with three subscales: Control (with four items from Abe & Ida's Control subscale and one from their Monitoring subscale); Metacognitive Knowledge (with four items from Abe & Ida's Metacognitive Knowledge subscale and one from their Control subscale); and Monitoring (three items from the originally hypothesized Monitoring subscale and two from the Metacognitive Knowledge subscale). While there is possible evidence for a three-factor structure underlying the scores of Japanese university students on the MAI, in contrast to the two- and eight-factor structure postulated in other contexts, the composition of these factors, and thus a clearer picture of just what constructs these factors are measuring, remains unsettled.

Given the important role that metacognition plays in language learning, in active learning and in the effectiveness of group-based learning approaches, and moreover, considering MEXT's increased emphasis on these latter two at all levels of the Japanese educational system, there is a need for the evidence-based measurement of learners' metacognitive awareness. Because of its solid grounding in metacognitive theory and its prominence in both the international and the Japanese literature, the MAI (see Appendix) might serve as one form of such measurement. However, as discussed above the underlying structure of the MAI remains in question. The present study, reporting on an examination of the psychometric properties of the MAI, represents an incremental step towards providing a secure empirical foundation for research in this area.

In this study, five models for the MAI were tested. Four of these come directly out of the literature, and were described in detail above: 1) Schraw and Dennison's (1994) originally hypothesized two-factor structure (Model 1 in the analysis below); 2) Abe and Ida's (2010) three-factor hierarchical structure (Model 2); 3) Niwa and Yamaji's (2017) three-factor structure (Model 4); and 4) Teo and Lee's (2012) three-factor structure (Model 5) The remaining model, Model 3, was a correlated, rather than hierarchal version of Abe and Ida's postulated model. This fifth model was examined in order to ascertain whether Abe and Ida's supposition of a hierarchical structure, which is unique among all the hypothesized structures in the literature, was more plausible than that of a simple correlation between the factors.

Methodology

The Methodology section of this study is reported in three sections below. First, the instrument that is the focus of this study, the MAI, is described. Following this, the data collection process and a characterization of the participants is provided. Finally, the procedures undertaken to analyze the data gathered are delineated.

Instrument

Schraw and Dennison's MAI (1994) is a 44-item instrument that consists of two hypothesized subscales. These were taken by its authors to measure two broad categories of respondents' metacognitive awareness; that is, the KC subscale (Items 3, 5, 7, 9, 10, 12, 13, 15, 17, 18, 20, 25, 26, 29, 30, 31, 32, 33, 39, 42, 45, 46, 51 and 52), and the RC subscale (Items 1, 2, 4, 6, 8, 11, 14, 16, 19, 21, 22, 23, 24, 27, 28, 34, 35, 36, 37, 38, 40, 41, 43, 44, 47, 48, 49 and 50). In the original study responses were recorded on a 100 mm line, semantically anchored with *false* at one endpoint and *true* at the other, and respondents were asked to make a slash at the point on the line that best indicated how true or false the item was of them. While this response system has been used in a few studies since Schraw and Dennison's (1994) original study (e.g., Magno, 2010), the majority of studies have employed a Likert-scale response system (e.g., Akin, et al., 2007; Sperling, Howard, Staley, & DuBois., 2004; Teo & Lee, 2012; Young & Fry, 2008). This is also true of studies done in Japan (e.g., Abe & Ida, 2010; Niwa & Yamaji, 2017). For this reason, a response system based on a semantically anchored Likert scale, ranging from 1, *not true of me at all*, to 5, *very true of me* (see Appendix for the MAI instrument), was adopted in this study. In the original study, respondents' scores on the MAI were reported as a composite for each subscale, but there was no composite for the entire instrument. In other words, the subscales were not hypothesized to represent constructs which fall under a superordinate construct, or put another way, there was no hierarchical model for the instrument in its original conception.

Participants and Data Collection Process

A total of 729 respondents took part in this study. These respondents were students at four universities located in Western Japan. Participation in the survey was completely voluntary and formed no part of the respondents' assessment in the class. Informed consent was obtained through the inclusion of a form at the beginning of each questionnaire. This form clearly stated in Japanese that those students who did not wish to take part in the survey could do so simply by not completing the questionnaire. There was no time limit specified by the administrators of the instrument, however, the time required for the students to complete the questionnaire was approximately 15 minutes.

From the total 729 records collected from the respondents, 38 were found to have missing data and were therefore removed from the dataset. The removal of these records was considered not to have had an effect on the overall properties of the dataset as there was no discernable systematic pattern to the missing responses, as determined by inspection by the author. The resultant 691 records form the basis for the analysis described in the following section. Among the respondents, there were 386 males and 289 females (7 respondents did not indicate their gender). 44.6% of the respondents were in their 1st, and 33.4 % in their 2nd year of study, and the age range of the participants was from 17 to 25.

Analytical Procedure

The data gathered from the participants (which in addition to the scores on the MAI, included their age, gender, school year, university and department) was initially entered into a database using Microsoft Access 2016. For the purpose of generating descriptive statistics and reliability estimates (Cronbach's alpha) the database was imported into IBM/Statistical Package for the Social Sciences (SPSS; Version 21). For the CFAs carried out on the scores in the data set, which constitute the primary analysis in this study, measurement models for each of the five different postulated structures of the MAI under consideration were laid out and tested using AMOS (Version 21).

The descriptive statistics serve as a useful summary of the overall characteristics of the data set. Moreover, the degree of skew and kurtosis reported for each of the items can provide indications of the degree of univariate normality in the scores. Following the determination of the degree of skew and kurtosis for each item, reliability estimates were calculated for each of the hypothesized subscales for each of the four reported versions of the MAI under consideration in this study. Finally, CFAs were undertaken on each of the structures described above.

In order to determine the degree of skew and kurtosis found in the scores, the skew and kurtosis values for each of the items was divided by the respective standard errors to calculate a critical ratio. To assess the extent to which the skew and kurtosis found in the scores differ from a normal distribution, two criteria were adopted—a strict criterion of less than (or equal to) 2.0, and more relaxed criterion of less than (or equal to) 3.0. These values correspond roughly to significance values of .01 and .05, respectively, that a critical ratio exceeding these values is due to actual non-normality in the sample, rather than chance (Hair, et al., 2104).

Cronbach's alpha was adopted as the estimate of reliability in this study. In addition to the value for alpha, confidence intervals (95%) were calculated as recommended by Fan and Thompson (2001). A minimum value of .70, following Nunnally and Bernstein (1994), was employed as the criteria for interpreting the reliability of each of the subscales examined in this study.

The purpose of CFA is to evaluate the hypothesis that a postulated structure for the instrument in question replicates the relationships found in a set of empirical data, in the form of responses to a survey instrument, to a reasonable degree. In order to test this hypothesis, the postulated structure is operationalised in a measurement model and compared to the structure found in the scores (Harrison & Valin, 2017), by examining the degree to which the estimated variance-covariance matrix of the model resembles the actual variance-covariance matrix found in the data set. The result of this comparison is evaluated using a range of measures, each providing information on the extent to which the relationships postulated in the model fit the relationships in the scores.

In evaluating the results of the CFAs conducted in this study five decision points were employed to make the determination of model fit following the recommendations of Brown (2015). One of the five decision points was the $\chi 2$ (chi-square) test statistic. As the name suggests, this is not an index, and is interpreted according to the rules associated with typical inferential statistics, although the null model is favored over the alternative model if model fit is desirable; that is, the result should be non-significant rather than significant at some pre-determined level for alpha. The other four decision points involve indices. Two of these indices, the root mean squared error of approximation (RMSEA) and the standardized root mean square residual (SRMR), are considered absolute fit indices; that is, they provide a direct measure of how well the model specified by the researcher reproduces the observed data (Hair, et. al., 2014). Comparative, or incremental, fit indices, on the other hand, assess the fit of the hypothesized model in relation to a baseline model. The Tucker-Lewis index (TLI), and the comparative fit index (CFI) were adopted in this case. Brown (2015) suggests taking into consideration the results of indices from each of these categories when evaluating the hypothesis that the postulated model fits the empirical data adequately. For this reason, the results of χ2 test statistic and the four the goodness-fit-indices were used in conjunction in order to determine the degree of fit for each model tested.

The χ2 test statistic tests whether the postulated model fits the sample exactly (Brown, 2015), with the resultant value being interpreted against a probability level, using either $p < .05$ or $p < .01$ as the decision criterion. In principle, the larger the value for the χ2 is, the worse the model fits the data. According to Brown (2015), a poor, that is to say a large, χ2 value means that the "model estimates do not sufficiently reproduce the sample variances and covariances (i.e., the model does not fit the data well)," (p. 81). The χ2 is sensitive to size of the sample, and thus will almost always indicate less than adequate fit with a large sample size (Harrington, 2009). The value of χ2 also tends to increase as the number of items on an instrument increases (Hair, et al., 2014), and thus for longer instruments the value of χ2 may falsely indicate poor model fit. It is the presence of these limitations which have prompted the widespread use of indices of model fit.

The RMSEA, an absolute fit index, is one of the indices employed to evaluate the models in this study. This index assesses the extent to which the postulated model fits the observed structure of the data, and it does so on a continuum, as do all of the indices, rather than exactly as with the χ2 (Brown, 2015). The χ2 is used to either accept or reject depending of the associated probability level and is not therefore on a continuum of interpretation. The RMSEA is less sensitive to sample size, and thus it overcomes one of the weaknesses of the χ2. It is, however, sensitive to the complexity of the model being tested, and purposefully so, with more complex models being penalized. This comports with the general scientific principle that the best model is always the simplest one. An additional reason for the use of the RMSEA is that confidence intervals can be calculated for values generated by the RMSEA (Byrne, 2016). This interval provides information concerning the degree of imprecision in the reported value of the RMSEA. A wide confidence interval indicates a large degree of imprecision, whereas a narrow interval indicates the opposite.

The second absolute fit index, the SRMR, can be seen as an average of the residuals, that is average of the differences in the correlations found in the dataset against those predicted by the model (Brown, 2015). As it is a standardized measure, its value ranges from 0 to 1, with higher values indicating a less adequate degrees of fit.

The two incremental fit indices employed in this study are the CFI and the TLI. These two are among the "best behaved" (Brown, 2015, p. 72) of the goodness-of-fit indices, meaning that they have been found to be relatively insensitive to factors such as sample size or model complexity.

For both indices, higher values indicate a greater degree of fit between the model and the data.

In this study, the threshold criteria suggested by Hu and Bentler (1999) for each of the above indices (RMSEA <.06; SRMR <.08; TLI and CFI >.95) have been adopted when assessing model fit according to the value returned for each index. These thresholds were empirically derived by Hu and Bentler using simulations; that is, datasets with known properties built into them, and from which the behaviors of the respective indices could be observed.

Results

The results section of this study consists of three sections. First, the properties of the sample and the distribution of the scores on each item, which include the means, the standard deviation, and skew and kurtosis for each item, are presented in a report of the descriptive statistics. In the second section, the reliability estimates (Cronbach's alpha and the associated confidence levels for alpha) are presented. The third section comprises the results of the CFAs on each of the five models tested, and represents the critical part of the analysis.

Descriptive Statistics

The mean, standard deviation, and skew and kurtosis for each of the 52 items on the MAI were calculated as a summary of the characteristics of the dataset. The item with the highest mean was Item 46 (4.24), while Item 36 had the lowest mean (2.36). The standard deviation among the items ranged from 0.773 (Item 03) to 1.232 (Item 37).

Values for the critical ratios for the skew and kurtosis of each item were calculated as measures of the degree of normality in the scores. In terms of skew, the critical ratio for 36.5% of the items (19 of 52) fell below the strict 2.0 threshold, while that for 17.3% (9 of 52) of the items was between 2.0 and 3.0. The critical ratio for the remaining 46.2% (24 of 52) of the items exceeded the 3.0 threshold. The values of the critical ratio for the kurtosis of each of the items were roughly parallel, with 25.0% of the items (13 of 51) of the items possessing a critical ratio of less than 2.0, 23.1% (12 of 52) of the items possessing a critical ratio between 2.0 and 3.0, and 51.9% (27 of 52) with a critical ratio exceeding the relaxed threshold of 3.0. From these results, we can see that there is a degree of univariate non-normality in this dataset. This is noteworthy as the distribution of the data can influence the behavior of the goodness-of-fit indices, in particular incremental fit indices (Hair, et al., 2014). Harrington (2009) notes that in extreme cases, non-normality can result in the underestimation of both the TLI and the CFI.

Because of the degree of non-normality extant in the data employed in this study, the generalizability of its conclusions need to be qualified. However, as Schraw and Dennison (1994) did not report the degree of non-normality found in their original study, it is difficult to determine whether the degree of non-normality in the data is similar to that of the original study. In addition, this previous under-reporting also makes it difficult to determine whether the degree of non-normality exhibited by the data in this study is characteristic only of this particular sample, or this particular adaptation of the MAI, or whether a similar degree of non-normality might be observed in datasets from other populations, making this an issue which would be population invariant.

Reliability Estimates

The reliability estimates (Cronbach's alpha) and the confidence intervals (95%) for each of the subscales from each of the five models tested in this study are shown in Table 1 below.

Table 1

Reliability Estimates, Confidence Intervals for Alpha (95%), Scale Means, and Scale Standard Deviations for Scores on the Subscales of the MAI

Model	**Construct**	**No. of Items**	**Cronbach's alpha**	**95% Confidence Intervals for Cronbach's alpha**	
				Lower Bound	Upper Bound
1	Knowledge of Cognition	25	.845	.828	.862
1	Regulation of Cognition	19	.846	.829	.862
2	Monitoring	11	.771	.745	.795
2	Control	9	.750	.721	.777
2	Metacognitive Knowledge	8	.721	.698	.752
4	Control	5	.606	.557	.650
4	Metacognitive Knowledge	5	.700	.663	.734
4	Monitoring	5	.540	.483	.592
5	Conditional Knowledge	9	.691	.655	.724
5	Strategic Knowledge	3	.442	.365	.510
5	Procedural Knowledge	9	.720	.688	.750

Model 1: Schraw & Dennison, 1994; Model 2: Abe & Ida, 2010;
Model 4: Niwa & Yamaji, 2017; Model 5: Teo & Lee, 2012

The alpha for both subscales from Schraw and Dennison's (1994) original study was rather high, suggesting that both subscales may possess an adequate degree of reliability in this dataset. These values, while lower, are similar to the values reported in the original study; that is, .91 for both subscales. The subscales extracted in Abe and Ida's EFA study also performed adequately, with all three exceeding the .70 threshold (discussed above in the Methodology section). Only the lower bound confidence interval for the third subscale, Metacognitive Knowledge, failed to surpass this threshold (this, however, was by a mere .002). The three subscales from Niwa and Yamaji (2017) did not perform as well when measured in this dataset. Only the second subscale, also termed Metacognitive Knowledge, achieved the .70 threshold. The lower bound confidence interval for this subscale, while below the threshold value, is relatively close to .70 and thus an argument could be made that this subscale exhibits a sufficient degree of reliability. This argument might be strengthened considering the findings of Cortina (1993) and Green, Lissitz and Muliak (1977), whose research showed that alpha tends to favor subscales with a larger number of items, and thus subscales with fewer items tend to produce lower alphas. A similar argument might be proffered to explain the poor performance of Niwa and Yamaji's two remaining subscales, however the alpha values for the subscales are rather far beneath the threshold in both cases, and even the upper bonds of the confidence intervals failed to reach an acceptable value. Finally, of the three subscales from Teo and Lee's (2012) study, the third subscale, Procedural Knowledge, surpassed the threshold, while the first subscale, Conditional Knowledge, missed the mark by only .009. The second subscale, however, performed quite poorly, with the lowest value by a wide margin among the eleven subscales listed in Table 1. Overall, however, there is a clear relationship between the number of items and the value for alpha in these results, and in fact the results may be considered an exemplar of the limitation of the index, pointed out by Green, Lissitz and Muliak. As such these results should be treated critically, and are methodologically subordinate to the results of the CFAs reported below.

Confirmatory Factor Analysis

As discussed above in the Methodology section, the purpose of CFA is to directly test the hypothesis that the relationships postulated in the model, i.e. between the items and their respective underlying constructs, resemble the actual relationships found in a set of empirical data to a reasonable degree. The five theoretical models for metacognitive awareness under consideration in this study were tested and the outcomes of these tests were evaluated using the $\chi 2$ test statistic in conjunction with the four goodness-of-fit indices, as discussed above. The value returned for each of these indices and for each model is shown in Table 2 below. In addition to these values, the value for Mardia's coefficient, as a measure of multivariate normality is also provided for each model. The value for this coefficient will not differ when the identical items appear in a model, even under a different specification. A value over 5.0 for this measure indicates multivariate non-normality in the dataset in question, and there will also be a tendency for the value to increase as the number of items rises.

Table 2
Comparison of Goodness-of-fit Indicators for the Measurement Models of the MAI

	TLI	**CFI**	**RMSEA**	**SRMR**	**Mardia's Coefficient**	**χ2**	**(p)**
	(>.95)	(>.95)	(<.06)	(<.08)	(< 5)		
Model 1 (Schraw & Dennison, 1994)	.658	.675	.062	.0685	86.043	3286.695	.000
Model 2 (Abe & Ida, 2010)	.757	.776	.064	.0723	64.496	1315.076	.000
Model 2 (error constrained) (Abe & Ida, 2010)	.751	.771	.064	.0727	64.496	1339.530	.000
Model 3 (Based on Abe & Ida, 2010)	.757	.776	.064	.0727	64.496	1339.530	.000
Model 4 (Niwa & Yamaji, 2017)	.792	.828	.070	.0645	41.997	378.605	.000
Model 5 (Teo & Lee, 2012)	.758	.786	.065	.0599	45.815	735.413	.000

*This model generated a negative variance in one of the error terms, and thus the solution is inadmissible. The results are presented here for the purposes of comparison.
TLI: Tucker-Lewis index; CFI: Comparative fit index; RMSEA: root mean squared error of approximation; SRMR: standardized root mean square residual; χ2: Chi-square test statistic.
Values in parentheses are Hu and Bentler's (1999) cut-off values.

Model 1: Two-Factor Model for the MAI (Schraw & Dennison, 1994)

This measurement model was constructed on the basis of the EFA results in Schraw and Dennison's (1994) original study. The model comprised two correlated factors with 990 distinct sample moments, 89 distinct parameters to be estimated, and therefore 901 degrees of freedom. These values met the criteria for overidentification. A model is considered overidentified when the number of parameters to be estimated is less than the number of distinct sample moments, that is, less than the number of variances and covariances of the observed variables in the model (Byrne, 2016).

The degree of multivariate normality was assessed using Mardia's coefficient, where a value exceeding 5.0 signifies a degree of multivariate non-normality. The value for Mardia's coefficient for this model was 86.043, indicating multivariate non-normality in the scores. The χ2 value for this model was 3286.695 with a probability level of .000. The results from the calculations of the fit indices were as follows (with Hu and Bentler's [1999] cut-off values given in parentheses): TLI .658 (>.95), CFI .675 (>.95); RMSEA .062 (<.06); SRMR .0685 (<.08). Taken in conjunction, the four goodness-of-fit indices and the χ2 indicate that the two-factor model hypothesized by Schraw and Dennison (1994) does not fit the data to a sufficient degree and thus should be rejected.

Model 2: Second-Order Model for the MAI (Abe & Ida, 2010)

The second model tested was derived from the results of Abe and Ida's (2010) EFA and subsequent CFA. As mentioned above, this model is unique in that Abe and Ida postulated a hierarchical structure, with a second-order factor, termed Metacognition, presumed to explain the degree of correlation found between the three factors. This model possessed 406 distinct sample moments, 59 distinct parameters to be estimated, and 347 degrees of freedom, and thus was overidentified. While the values for Mardia's coefficient, the χ2 and the other goodness-of-fit indices for this model are reported in Table 2, it should be noted that this model produced an inadmissible solution. The source of this inadmissibility was the appearance of a negative variance in one of the error terms in the model as initially specified. More specifically, one of the first order factors in the model was calculated to have an error of less than zero. Conversely, this means that the degree of correlation between this factor and the postulated second-order factor was greater than 1, which is an unreasonable value. This type of situation, where the postulated model generates an illogical value for one its parameters, is known as a Heywood case. The consequences of this result for the model in question and its viability are considered further in the Discussion section below, however, one suggested solution for the appearance of a negative variance in one of the error terms is to constrain the variance of the error term in the model to a value close to 0 (Chen, Bollen, Paxton, Curran & Kirby, 2001). The model was thus re-specified, with the variance of the error term constrained to 0.001. This constrained error model produced an admissible solution. This model had 406 distinct sample moments, 58 distinct parameters to be estimated, and 348 degrees of freedom, and thus was also overidentified.

Mardia's coefficient for this model was 64.496, indicating multivariate non-normality in the data, and would be expected to be the same value as for the Heywood-case model, because the items are the same. The results for this model were as follows (Hu and Bentler's [1999] cut-off values in parentheses): TLI .751 (>.95), CFI .771 (>.95); RMSEA .064 (<.06); SRMSR .0727 (<.08). The χ2 value was 1339.530 with a probability level of .000. These results do not differ greatly from those of the unconstrained model, suggesting that constraining the variance of the error term did not alter the relationships in the model to an appreciable degree. The results indicate that even in its constrained form this three-factor hierarchical model fails to reproduce the structure found in the data to an adequate degree and therefore should also be rejected.

Model 3: Three-Factor Rival Model for the MAI

As the model discussed above possessed a hierarchical structure, an additional measurement model, comprised of the same three factors, but allowed to correlate rather than being placed as subordinate to a further construct, was tested as well. This model met the criterion for overidentification, with 406 distinct sample moments, 59 distinct parameters to be estimated, and 347 degrees of freedom.

As shown in Table 2 above, the values for Mardia's coefficient, the χ2, CFI, TLI and RMSEA were identical to those for the unconstrained hierarchical model (Model 2), with only the SRMR differing to a small degree. As with the hierarchical model, these results strongly suggest that the structure of the correlated model does not sufficiently correspond to the structure underlying the scores in the dataset. For this reason, the correlated model should also be rejected. The implications of the performance of this model and its relationship with the hierarchical model will be examined in the discussion that follows.

Model 4: Three-Factor Model for the MAI (Niwa & Yamaji, 2017)

Measurement Model 4 was formulated on the basis of the results of the EFAs carried out by Niwa and Yamaji (2017). This model comprised three correlated factors, with 120 distinct sample moments, 33 distinct parameters to be estimated, and 87 degrees of freedom, and thus was overidentified.

Multivariate non-normality in the data was indicated by the value for Mardia's coefficient for this model, 41.997. The value of χ2 was 378.605 with a probability level of .000. The results of the fit indices for this model were as follows (Hu and Bentler's [1999] values in parentheses): TLI .729 (>.95), CFI .828 (>.95); RMSEA .070 (<.06); SRMSR .0645 (<.08). The value for the χ2 for this model, while still indicating that the model failed to replicate the relationships in the data sufficiently, was the lowest of all the models (though this may simply be a result of the reduced number of items in the model). Nonetheless, all of the values suggest that this model does not fit the data sufficiently, and therefore it should also be rejected.

Model 5: Three-Factor Model for the MAI (Teo & Lee, 2012)

The final model tested in this study, Model 5, was derived from the results of the EFA and CFAs described by Teo and Lee (2012). This model also comprised three correlated factors, and met the criterion for overidentification, with 231 distinct sample moments, 45 distinct parameters to be estimated, and 186 degrees of freedom.

The results of the goodness-of-fit indices for this model were as follows, (Hu and Bentler's [1999] cut-off values given in parentheses): TLI .758 (>.95), CFI .786 (>.95); RMSEA .065 (<.06); SRMSR .0599 (<.08). The χ2 value was 735.413 with a probability level of .000. Mardia's coefficient for this model was 45.815, indicating multivariate non-normality in the data. As with the previous models, the results taken in conjunction, indicate that the model hypothesized by Teo and Lee fails to fit the structure of the scores in the dataset sufficiently and therefore this model should also be rejected.

Discussion

The dimensionality of scores produced by the MAI in the Japanese EFL context are the focus of the results reported above. The MAI is a widely-used instrument for measuring respondents' degree of metacognitive awareness in both the international and Japanese context. Despite its widespread use, the underlying structure of the MAI remains contentious, with previous research indicating two, three or even the possibility of eight constructs underlying the items on the instrument. The aim of this paper was to examine the plausibility of several hypothesized models using CFA in order to determine which, if any, were valid explanations of scores generated in the Japanese university student population. If a viable model were to be found, the MAI could serve as a useful tool in exploring the role of metacognition in second language acquisition and group-based learning approaches, as well as other areas of interest where metacognition plays a significant role, such as active learning.

Disappointingly, however, the results from direct tests employing CFA as the principal method of analysis indicated that the five theoretical structures evaluated using measurement models exhibited an insufficient degree of fit with the structure underlying the empirical dataset gathered for this study. While these results may be unsatisfying in terms of the positive evidence required to recommend a plausible and evidence-based model for actual use by the researcher and practitioner, the negative evidence reported here does have the positive function of providing an empirically-grounded rationale for explaining why use of the instrument, by the researcher or practitioner, under currently advocated models is not recommended, and in helping to address future theorization and measurement of this class of constructs.

For one thing, it may be the case that a stable model for the structure underlying the MAI has yet to be found. As noted in the discussion above, a number of possible solutions have been reported in previous research. In their original study, Schraw and Dennison (1994) postulated eight subcomponents of metacognitive awareness. However, their initial EFA resulted in six factors, whose content was not interpretable in terms of theory. When a two-factor solution was imposed on the data, the two factors comprised primally items from the two major theoretical categories of knowledge of cognition and regulation of cognition, respectively. It must be noted that even in this imposed two-factor solution, eleven items postulated to belong to the regulation of cognition construct loaded onto the factor composed primarily of items related to the knowledge of cognition construct. Thus, even in its original form, there may be an empirical departure from the theory informing the supposed latent structure of items comprising the MAI.

Moreover, even in studies where solutions with a similar number of latent factors were proposed, such as in Abe and Ida (2010) and Teo and Lee (2012), who both postulated three constructs underlying the scores, the content of these factors differed to a large extent, leading to variations in the interpretation of the meaning of the factors and their underlying constructs. This fact, the content differences in an equivalent number of latent factors, holds true even in research using the same version of the MAI in similar populations, such as in Abe and Ida (2010) and Niwa and Yamaji (2017), both of which involved the Japanese undergraduate population. While the version of the MAI proposed by Abe and Ida was employed in Niwa and Yamaji's study, the results of their EFA showed a number of items loading on factors different from those in Abe and Ida's study.

Pintrich, Wolters and Baxter (2000) propose one source for this lack of stability in the structure of the MAI. Their paper examines the construct validity of both the MAI and the Index of Reading Awareness (Jacobs & Paris, 1987). Their review of the literature found that there may exist "theory-data mismatch" (p. 63) in both of these instruments. That is to say, metacognitive theories informing the instruments predicted a greater number of constructs than were observed in the empirical data generated by these instruments.

In their explanation of this mismatch, Pintrich et al. refer to the concept of *grain-size* propounded by Howard-Rose and Winnie (1993). This concept is concerned with the resolution—the degree of detail measurable, or the smallest interval discernable—of an instrument. According to this form of explanation, our theories regarding cognitive constructs, in this case metacognition and metacognitive awareness, suggest "small grain-size components, or relatively fine distinctions," (Pintrich, et al., 2000, p. 63). However, instruments which are grounded in these theories, such as the MAI, may not possess the necessary resolution to bring these fine distinctions into clear and operational focus.

This theory-data mismatch may go beyond questions concerning the instrument as a whole and may impact the content of individual items themselves. It may be difficult for respondents to make fine distinctions between some items on the MAI with the same degree of skill as an expert in the field, such as between items concerned with a knowledge of procedures and those concerned with putting the procedures into practice. Examples of such difficulties can be found in Schraw and Dennison's (1994) original study, where Item 14, *I have a specific purpose for each strategy I use*, and hypothesized to belong to the procedural knowledge subcomponent of knowledge of cognition, loaded on the factor comprising primarily regulation of cognition items. Another example would be where Item 42, *I read instructions carefully before I begin a task*, and Item 45, *I organize my time to best accomplish my goals*, are supposed to belong to the planning subcomponent of regulation of cognition, but loaded on the knowledge of cognition factor.

Further examples of this in the Japanese population, the population which is the primary focus of this study, can be found in the differing results in the work of Abe and Ida (2010) and Niwa and Yamaji (2017). Item 52, *I stop and reread when I get confused*, is an item on Abe and Ida's Control subscale, however, in Niwa and Yamaji's model this item shifts to their Metacognitive Knowledge subscale. This shift may be explained if one were to suppose that respondents in Niwa and Yamaji's study considered this item as more of general maxim to be followed when studying, and thus a constituent of one's knowledge of cognition, rather than as a specific procedure or strategy that one employs, as Abe and Ida's respondents may have taken this item to imply.

This brings us back to the premise of Pintrich, at al. (2000). It may be that experts, through their theoretical conceptions, can discern a fine-grained, or high-resolution, structure for metacognition (and possibly other psychological constructs), but neither our instruments nor the typical respondent are able to match this degree of precision, and thus they tend to perform at lower resolutions. As Pintrich et al. note, it remains a question for future research as to whether our instruments need to be more finely tuned in order to detect these fine grains, or if our theoretical conceptions and the instruments grounded in them should be refashioned to mirror the coarse grains coming out of empirical studies in this area.

A second issue that the results of this study may help to address is more closely related to the use of the MAI in the Japanese context, and is a question related to the structure of Abe and Ida's (2010) version of the MAI. To the best of this author's knowledge, this version of the MAI is the most widely-used in the Japanese context, and thus questions about its structure could have an impact on the findings of previous studies that used this version of the MAI.

As noted above, the structure hypothesized by Abe and Ida (2010) is unique in that it postulates a hierarchical structure for the instrument. Rather than three correlated factors, the structure suggested by Abe and Ida supposes a second-order factor, termed Metacognition, to explain the relationships among the three subscales on the instrument. This second-order factor was proposed on the basis of the correlation found between the three factors, which Abe and Ida (2010) characterized as moderate. The correlations between the three-factors reported in their study were between .401 and .562. The higher of these values does not differ greatly from the degree of correlation (r = .54) found in Schraw and Dennison's (1994) original study. However, Schraw and Dennison did not propose the existence of a second-order factor to explain the degree of correlation which they had observed.

In order to investigate the question of whether a hierarchical or correlated structure better replicated the structure of the empirical data, measurement models for both of these structures were tested in this study (Model 2 and Model 3 in Table 2). As discussed above in the results section, the values for the goodness-of-fit indices resulting from both models were quite similar. This suggests that the supposition of a second-order factor does not explain the underlying structure of the data to a greater degree than the simpler, correlated structure.

Moreover, the hierarchical model generated a Heywood case, in this instance a negative error variance. Only when this variance was constrained did this model produce an admissible solution. Such negative error variances may be due to a number of issues, e.g., small sample size, non-normality or outliers (although the restricted range of the Likert scale probably discounts the outlier issue), multicollinearity or model misspecification (Brown, 2015). The sample size in this study was 691, which exceeds the recommendation of Hair et al. (2014) for a ratio of 10 responses for each indicator, and so this is probably not a source of explanation for the problem. As mentioned above, a degree of univariate, as well as multivariate, non-normality was found in the dataset, and thus non-normality cannot be completely ruled out as the source of the inadmissible solution. That being the case, only the hierarchical model generated a Heywood case, while the solutions of all the other models tested were admissible, without any alterations.

Negative error variance may also often be a sign of model misspecification (Byrne, 2016), and it may the case that by postulating the second order factor, Abe and Ida have created a model which is misspecified to some extent. However, it may also be the case that by postulating this second-order factor they have engendered an instance of multicollinearity in their model. Byrne (2016) characterizes multicollinearity as two or more variables being so highly correlated that they essentially represent the same construct. The implication of a negative error variance is the existence of a degree of correlation greater than 1. In Model 2 (when the error variance was left unconstrained), such a correlation (r = 1.19) occurred between the first-order factor, Control, and the second-order factor, Metacognition. Even in the constrained version of Model 2 a perfect degree of correlation (r = 1.00) was found between these two factors in this study. The correlation between the first-order Control factor and the second-order Metacognition factor in Abe and Ida's (2010) study was also exceptionally high (r = .96). The fact that the correlation between these two factors in all three models is so remarkably similar argues against issues of sample size or non-normality in this study's dataset as the cause of the Heywood case, and represents evidence for either the multicollinearity of these two factors, or that the postulation of the second-order factor is a possible case of misspecification in Abe and Ida's model.

Brown (2015) suggests that when two factors correlate to a high degree (a suggested value of .85), questions can arise as to the discriminant validity of the two factors, and thus it might be possible to combine these two factors to achieve a more parsimonious model. While Brown's suggestion is directly addressing the issue of a high correlation between first-order factors, a similar logic might be applied to the postulated second-order factor in Abe and Ida's model. The evidence in the results of this study may not positively refute the possibility of this second-order factor, as both models displayed inadequate fit and were rejected. They also do not, however, provide it with any supporting evidence, in that it performed no better, and in some ways worse, than the correlated model, and for these reasons, as well as the principle of parsimony, it might well be the case that Abe and Ida's model requires revision in its present form.

Conclusion

Metacognition and metacognitive awareness play an arguably important role in collaboration, language learning and other areas of study. To aid in the expansion of research in this area, there is a need for evidence-based measurement of learners' metacognitive awareness. The MAI, because of its popularity and strong grounding in metacognitive theory, was thought to possess the potential to provide such measurement.

However, the contribution of the study reported here has to come in the form of negative evidence for the models thus far hypothesized for the MAI. This remains an important contribution nonetheless. Kline (2011, citing Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007), in his discussion of the role of hypothesis testing, remarks that,

> the real goal is to test a theory by specifying a model that represents predictions of that theory among plausible constructs measured with the appropriate indicators. If such a model does not ultimately fit the data, then this outcome is interesting because there is value in reporting models that challenge or debunk theories. (p. 189)

In the sense outlined in this quote, the results of the study on the MAI, and reported above, are interesting and valuable. They have shed light on the broader issue of discrepancies in the theoretical conception and empirical expression of metacognitive awareness when being measured through the MAI, as well as exposing a more specific methodological issue with respect to the possibility of misspecification in Abe and Ida's (2010) widely-used model. The data and results of this study could be of use in illuminating a path for possible revisions that might be undertaken with respect to Abe and Ide's model for the MAI in the Japanese EFL context. Once such changes have been made, the revised instrument could then be tested against a new dataset, and the fit of the model re-examined.

Finally, it must be noted that one limitation of this study is that it deals with only a single sample, and a sample that is one of convenience rather than a true representative sample of the target population; which is not unusual given the practical constraints in obtaining truly representative samples, but is a limitation nonetheless. For this reason, the generalizability of the findings of this study to the target population should be treated with appropriate levels of critical caution, and further research using other samples from the same population will be of assistance in addressing this limitation as part of the normal process of public research

References

阿部真美子 [Abe Mamiko] & 井田政則 [Ida Masanori]. (2010). 成人用メタ認知尺度の作成の試み：Metacognitive Awareness Inventory を用いて [An attempt to construct the adults' metacognition scale: Based on Metacognitive Awareness Inventory]. 立正大学心理学部 [*Journal of the Psychology Department of Rissho University*], *1*, 23-34.

Anderson, N. (2002). The role of metacognition in second language teaching and learning. (Report No. EDO-FL-01-1). Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED463659).

Anderson, N. J. (2008). Metacognition and good language learners. In C. Griffiths (Ed.), *Lessons from good language learners*, (pp. 99-109). Cambridge, UK: Cambridge University Press.

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*(2), 671-678.

青野健治 [Aono Kenji] & 西野泰代 [Nishino Yasuyo]. (2016). 理学療法士を目指す学生の学業成績を規定する要因についての検討 [Determinants of academic performance among physical therapy students]. 広島修大論集 [*Studies in the Humanities and Sciences, Hiroshima Shodo University*], *56*, 83-96.

Artz, A. F., & Armour-Thomas, E. (1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction, 9*(2), 137-175.

Brown, A., & Palincsar, A. (1989). Guided cooperative learning and individual knowledge acquisition. In L.B. Resnick (Ed.), *Knowing, learning and instruction. Essays in honor of Robert Glaser,* (pp. 393-451). Hillsdale, NJ: Erlbaum.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY.: Guilford Press.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York, NY: Routledge.

Cantwell, R. H., & Andrews, B. (2002). Cognitive and psychological factors underlying secondary school students' feelings towards group work. *Educational Psychology, 22*(1), 75-91.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29*(4), 468-508.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*(4), 827-838.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis: Pearson new international edition*. Essex, UK: Pearson Education Limited.

Harrington, D. (2009). *Confirmatory factor analysis*. New York, NY.: Oxford University Press.

Harrison, G. M., & Vallin, L. M. (2018). Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning, 13*(1), 15-38.

Haukås, Å. (2018). Metacognition in language learning and teaching: An overview. In A. Haukås, C. Bjørke & M. Dypedahl (Eds.), *Metacognition in Language Learning and Teaching*, (pp. 25-44). New York, NY: Routledge.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three-Testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841-850.

Howard-Rose, D., & Winne, P. H. (1993). Measuring component and sets of cognitive processes in self-regulated learning. *Journal of Educational Psychology, 85*(4), 591-604.

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), 1-55.

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22*(3-4), 255-278.

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

Kondo, M., Ishikawa, Y., Smith, C., Sakamoto, K., Shimomura, H., & Wada, N. (2012). Mobile assisted language learning in university EFL courses in Japan: Developing attitudes and skills for self-regulated learning. *ReCALL, 24*(2), 169-187.

Lee, C. B., Teo, T., & Bergin, D. (2009). Children's use of metacognition in solving everyday problems: An initial study from an Asian context. *The Australian Educational Researcher, 36*(3), 89-102.

Lumma-Sellenthin, A. (2012). Medical students' attitudes towards group and self-regulated learning. *International Journal of Medical Education, 3*, 46-56.

Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and Learning, 5*(2), 137-156.

Ministry of Education, Culture, Sports, Science and Technology. (2009a). 中学校学習指導要領「外国語」英訳版（仮訳）[Junior high school *Course of Study* (Foreign languages) English version (Provisional translation)]. Retrieved July 6th, 2018 from: http://www. mext. go.jp/component/english/__icsFiles/afieldfile/2011/03/17/ 1303755_013.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2009b). 中学校学習指導要領解説 [Explanatory comments for the junior high school *Course of Study*]. Retrieved July 20th, 2018 from:http://www.mext.go.jp/component/a_menu/ education/ micro_detail/__icsFiles/ afieldfile/2011/01/05/1234912_010.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2012). 新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～（答申）[Towards a qualitative transformation of university education for building a new future: Universities fostering lifelong learning and the ability to think independently and proactively (Report)]. Retrieved July 8th, 2018 from http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/ afieldfile/2012/10/04/1325048_1.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2013). Measures based on the four basic policy directions. In *The second basic plan for the promotion of education (Provisional translation)*, (Part 2, Section 1). Retrieved July 5th, 2018, from http://www. mext.go.jp/en/policy/education/lawandplan/title01/detail01/ sdetail01/1373805.htm

Ministry of Education, Culture, Sports, Science and Technology. (2014a). On integrated reforms in high school and university education and university entrance examination aimed at realizing a high school and university articulation system appropriate for a new era. Retrieved July 7th, 2018 from http://www.mext.go.jp/component/English/__icsFiles/afieldfile/2015/03/31/1353908_1.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2014b). Report on the future improvement and enhancement of English education (Outline): Five recommendations on the English education reform plan responding to the rapid globalization. Retrieved July 7th, 2018 from http://www.mext.go.jp/en/news/ topics/detail/1372625.htm

Ministry of Education, Culture, Sports, Science and Technology. (2016). Summary of report: Towards a qualitative transformation of university education for building a new future: Universities fostering lifelong learning and the ability to think independently and proactively. Retrieved July 7th, 2018 from http://www.mext. go.jp/en/publication/report/title01/detail01/__icsFiles/afieldfile/2016/12/06/1380275_001.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2017a). 中学校学習指導要領解説 [Explanatory comments for the junior high school *Course of Study*]. Retrieved July 20th, 2018, from http://www.mext.go.jp/component/ a_menu/ education/micro_detail/__icsFiles/afieldfi le/2018/05/07/1387018_10_1.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2017b). 高等学校学習指導要領 [Explanatory comments for the high school *Course of Study*]. Retrieved July 20th, 2018, from: http://www. mext.go.jp/component/a_menu/education / micro_detail/__icsFiles/afieldfile/2018/07/11/1384661_6_1_2.pdf

Molenaar, I., Sleegers, P., & van Boxtel, C. (2014). Metacognitive scaffolding during collaborative learning: A promising combination. *Metacognition and Learning, 9*(3), 309-332.

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*(1), 177-195.

丹羽量久 [Niwa Kazuhisa] & 山地弘起 [Yamaji Hiroki]. (2017). 初年次学生のメタ認知の測定 [Measurement of metacognitive awareness in first-year students]. 長崎大学大学教育イノベーションセンター紀要 [*Journal of the Center for Educational Innovation Nagasaki University*], *8*, 45-50.

Nunnally, J. C., & Bernstein, I. H. (1994). The assessment of reliability. *Psychometric Theory, 3*, 248-292.

Pintrich, P., Wolters, C. & Baxter, G (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. Impara (Eds), *Issues in the measurement of metacognition*, (pp. 43-98). Lincoln, NE: Buros Institute of Mental Measurements.

Raes, A., Schellens, T., De Wever, B., & Vanderhoven, E. (2012). Scaffolding information problem solving in web-based collaborative inquiry learning. *Computers & Education, 59*(1), 82-94.

Raoofi, S., Chan, S. H., Mukundan, J., & Rashid, S. M. (2014). Metacognition and second/foreign language learning. *English Language Teaching, 7*(1), 36-49.

齋藤ひとみ . [Saito Hitomi]. (2016). 課題遂行時間の見積もりと先延ばし行動および先延ばし意識との関係 [Investigation on the relationship among task time estimation, procrastination behavior, and awareness of procrastination]. 愛知教育大学研究報告 . 教育科学編 [*Bulletin of Aichi University of Education, Educational Sciences*] *65*, 181-186.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460-475.

Smith, J. M., & Mancy, R. (2018). Exploring the relationship between metacognitive and collaborative talk during group mathematical problem-solving–what do we mean by collaborative metacognition? *Research in Mathematics Education, 20*(1), 14-36.

Sperling, R. A., Howard, B. C., Miller, L. A., & Murphy, C. (2002). Measures of children's knowledge and regulation of cognition. *Contemporary Educational Psychology, 27*(1), 51-79.

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *Educational Research and Evaluation, 10*(2), 117-139.

Splichal, J. M., Oshima, J., & Oshima, R. (2018). Regulation of collaboration in project-based learning mediated by CSCL scripting reflection. *Computers & Education, 125*, 132-145.

Sungur, S. (2007). Contribution of motivational beliefs and metacognition to students' performance under consequential and nonconsequential test conditions. *Educational Research and Evaluation,* *13*(2), 127-142.

Teo, T., & Lee, C. (2012). Assessing the factorial validity of the Metacognitive Awareness Inventory (MAI) in an Asian country: A confirmatory factor analysis. *International Journal of Educational and Psychological Assessment, 10*(2), 92-103.

Victori, M. (1999). An analysis of writing knowledge in EFL composing: A case study of two effective and two less effective writers. *System, 27*(4), 537-555.

Webb, N. M. (2009). The teacher's role in promoting collaborative dialogue in the classroom. *British Journal of Educational Psychology, 79*(1), 1-28.

Xethakis, L. (2017). Creating conditions for collaborative learning in the language classroom. In G. Brooks (Ed.), *The 2016 PanSIG Journal*, (pp. 351-359). Tokyo, Japan: JALT.

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.

Zhang, Y. (2010). Cooperative language learning and foreign language learning and teaching. *Journal of Language Teaching and Research, 1*(1), 81-83.

Appendix – Metacognitive Awareness Inventory (Schraw & Dennison, 1994)

1. I ask myself periodically if I am meeting my goals.
2. I consider several alternatives to a problem before I answer.
3. I try to use strategies that have worked in the past.
4. I pace myself while learning in order to have enough time.
5. I understand my intellectual strengths and weaknesses.
6. I think about what I really need to learn before I begin a task.
7. I know how well I did once I finish a test.
8. I set specific goals before I begin a task.
9. I slow down when I encounter important information.
10. I know what kind of information is most important to learn.
11. I ask myself if I have considered all options when solving a problem.
12. I am good at organizing information.
13. I consciously focus my attention on important information.
14. I have a specific purpose for each strategy I use.
15. I learn best when I know something about the topic.
16. I know what the teacher expects me to learn.
17. I am good at remembering information.
18. I use different learning strategies depending on the situation.
19. I ask myself if there was an easier way to do things after I finish a task.
20. I have control over how well I learn.
21. I periodically review to help me understand important relationships.
22. I ask myself questions about the material before I begin.

23. I think of several ways to solve a problem and choose the best one.
24. I summarize what I've learned after I finish.
25. I ask others for help when I don't understand something.
26. I can motivate myself to learn when I need to.
27. I am aware of what strategies I use when I study.
28. I find myself analyzing the usefulness of strategies while I study.
29. I use my intellectual strengths to compensate for my weaknesses.
30. I focus on the meaning and significance of new information.
31. I create my own examples to make information more meaningful.
32. I am a good judge of how well I understand something.
33. I find myself using helpful learning strategies automatically.
34. I find myself pausing regularly to check my comprehension.
35. I know when each strategy I use will be most effective.
36. I ask myself how well I accomplished my goals once I'm finished.
37. I draw pictures or diagrams to help me understand while learning.
38. I ask myself if I have considered all options after I solve a problem.
39. I try to translate new information into my own words.
40. I change strategies when I fail to understand.
41. I use the organizational structure of the text to help me learn.
42. I read instructions carefully before I begin a task.
43. I ask myself if what I'm reading is related to what I already know.
44. I re-evaluate my assumptions when I get confused.
45. I organize my time to best accomplish my goals.
46. I learn more when I am interested in the topic.
47. I try to break studying down into smaller steps.
48. I focus on overall meaning rather than specifics.
49. I ask myself questions about how well I am doing while I am learning something new.
50. I ask myself if I learned as much as I could have once I finish a task.
51. I stop and go back over new information that is not clear.
52. I stop and reread when I get confused.

日本での第二言語獲得場面における日本語版メタ認知尺度の心理測定

セタキス・ラリー

本研究では、第二言語獲得場面におけるメタ認知尺度（MAI;Schraw & Dennison, 1994）と、先行研究において最も使用されていた３つのバージョンのメタ認知尺度における心理測定特性について報告する。西日本に所在する４大学の学生729名から回答を得、これを分析対象とした。調査項目の正規性が確認でき、４バージョン全てにおいて下位尺度の信頼性をクロンバックα係数により確認した。確証的因子分析を用い、MAIの以下４つのバージョンモデルの構成概念妥当性を検証した、１）Schraw・Dennison（1994）による仮説による相関のある２因子構造；２）阿部・井田（2010）による階層のある３因子構造；３）丹羽・山地（2017）相関のある３因子構造；４）Teo・Lee（2012）による相関のある３因子構造。本調査結果の結果を用い、４つのモデルとの適合度を分析したが、本調査における構造との適合度は不良であった。本研究において、MAIのスコアから２因子および３因子構造を支持しない結果を得たことは、今後のメタ認知構造の理論化において有益な知見となり得る。