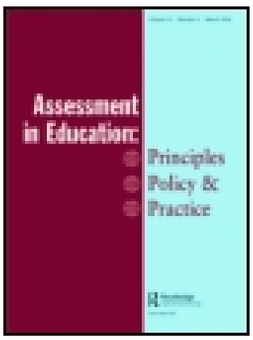

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Cross-cultural comparability and validity of metacognitive knowledge in reading in PISA 2009: a comparison of two scoring methods

Ji Zhou , Jia He & Dominique Lafontaine

To cite this article: Ji Zhou , Jia He & Dominique Lafontaine (2020): Cross-cultural comparability and validity of metacognitive knowledge in reading in PISA 2009: a comparison of two scoring methods, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2020.1828820

To link to this article: https://doi.org/10.1080/0969594X.2020.1828820

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
https://doi.org/10.1080/0969594X.2020.1828820

Cross-cultural comparability and validity of metacognitive knowledge in reading in PISA 2009: a comparison of two scoring methods

Ji Zhou [a], Jia He [a,b] and Dominique Lafontaine [c]

[a]Leibniz-Institut Für Bildungsforschung Und Bildungsinformation (DIPF), Frankfurt am Main, Germany; [b]Department of Methodology and Statistics, Leibniz-Institut Für Bildungsforschung Und Bildungsinformation (DIPF), Tilburg University, Tilburg, Netherlands; [c]Department of Education Science, University of Liège, Liège, Belgium

ABSTRACT

Accurate measurement of metacognitive knowledge in reading is important. Different instruments and scoring methods have been proposed but not systematically compared for their measurement comparability across cultures and validity. Given student data from 34 OECD countries in the Programme for International Student Assessment (PISA) in 2009, we compared two scoring methods for metacognitive knowledge in reading based on pair-wise comparisons of strategies and with conventional Likert-scale responses of selected items. Metacognitive knowledge scored with conventional Likert-scale responses demonstrated higher cross-cultural comparability than the pair-wise comparison method. Linked with reading competence, motivation and control strategy in reading, scores from the two scoring methods showed differential criterion validity, possibly related to the types of tasks (understanding and remembering versus summarising), item content (complexity and discrimination between preferred strategies in reading) and common method variance (e.g., individuals' stable response style in rating scales). Theoretical and methodological implications are discussed..

ARTICLE HISTORY
Received 9 January 2020
Accepted 8 September 2020

KEYWORDS
Metacognitive knowledge; reading; cross-cultural comparability; validity; Programme for International Student Assessment (PISA)

Metacognitive knowledge in reading

Definition

Metacognition was referred by Flavell as the active monitoring and consequent regulation of learning process in service of concrete goals (Flavell, 1976). The Organisation for Economic Co-Operation and Development (OECD) endorsed a similar definition of metacognition for the Programme for International Student Assessment (PISA) 2009: in the framework for reading (OECD, 2010, p. 72), metacognition in reading is defined as 'the awareness of and ability to use a variety of appropriate strategies when processing texts in a goal-oriented manner.'

Most of the theoretical models of metacognition distinguish between metacognitive knowledge on the one hand, and control/monitoring/regulation on the other hand. Flavell

CONTACT Ji Zhou ji.zhou@outlook.de Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF), Rostocker Street. 6, Frankfurt am Main 60323, Germany

(1976) defines metacognitive knowledge as knowledge about persons, tasks, and strategies. The knowledge about strategies could be further categorised into declarative, procedural, and conditional strategy knowledge (Flavell, 1976; Paris et al., 1983).

Assessing metacognitive knowledge

There is ample evidence that metacognitive knowledge is associated with higher reading proficiency, it is thus important to develop and strengthen it (Cubukcu, 2008; Taraban et al., 2004). Accurate measurement of metacognitive knowledge is crucial in informing researchers and policy makers about important components and prerequisites of the metacognition, also it is vital in monitoring students' progress in reading and constructing evidence-based intervention. As metacognition is related to the internal processes of non-overt behaviours covering a broad range of strategies and their uses in different contexts, its measurement is challenging (Allen & Armour-Thomas, 1993). Various direct and indirect (proxy) measures of metacognitive knowledge have been developed such as the Index of Reading Awareness by Jacobs and Paris (1987), and validations were carried out to inform understanding of the construct and functions of metacognitive knowledge (e.g., Akturk & Sahin, 2011; Jacobs & Paris, 1987; Love et al., 2019). It remains challenging to assess a broad range of metacognitive knowledge with brief measures, and to develop instruments which are reliable, show robust correlations with achievement and have a clear benchmark of evaluation. In comparative large-scale assessments, another requirement in terms of validity is that instruments need to be comparable across countries.

Metacognitive knowledge in reading in PISA 2009

In PISA 2009, a measure of metacognitive knowledge was implemented in the student questionnaire for the first time. In the PISA assessment, the conditional and relational strategy knowledge was considered the key component when students had to decide on more appropriate strategies (Artelt & Schneider, 2015; OECD, 2010). The approach was a scenario-based test: Two reading scenarios were presented and students had to rate the usefulness of possible strategies for each reading scenario. The measure was unique in the respect that in parallel to students' ratings, expert ratings of the usefulness of these strategies were collected and served as a benchmark against which students' ratings were compared. The relative preferences of pairs of strategies agreed upon by at least 80% of experts were selected to guide the scoring of students' data. When students' comparative judgements on these pairs were in line with the experts' ratings, they received a score of 1 and otherwise a score of 0, irrespective of their exact ratings on the Likert scale.

This scoring method has the advantage of including a clear benchmark (Artelt & Schneider, 2015). However, the original metrics from students' raw responses get lost when transforming the raw Likert-scale data into ranking data. For example, two students would receive the same score if both of them rated strategy B as being more useful than A, although one might consider both strategies to be 'useful' (e.g., rating of 6 and 5 for B and A, respectively) whereas the other considers both to be 'not useful' (rating of 2 and 1 for B and A, respectively). In this case, the absolute differences across students cannot be identified. To avoid information loss, metacognitive knowledge can also be measured with students' raw responses. The total score of this measure for each student is

the sum (or mean rating) of all item responses, or the factor score of a latent factor measured by these items. This alternative scoring approach is closer to traditional measurement of metacognitive strategies and similar to many measures in educational research, such as Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich et al., 1991) and Learning and Study Strategies Inventory (LASSI, Weinstein et al., 1987).

Despite the fact that the PISA instrument has been widely used in research (Mak et al., 2017; Säälik, 2015), there is a lack of empirical evidence on the cross-cultural comparability and criterion validity of the PISA 2009 metacognition scenarios. Two empirical studies reported on the measure of metacognitive knowledge scored based on the pair-wise comparison method. Artelt and Schneider (2015) tested the relationships across metacognitive knowledge, general control and reading achievement, and whether these relationships were comparable across countries. They found moderate to high correlations between metacognitive knowledge and reading competence (r = 0.48 on average across OECD countries), but lower correlations between metacognitive knowledge and control strategy use (r = 0.25 on average across OECD countries). Hence, Artelt and Schneider (2015) concluded on the cross-country generalisability of the role of metacognitive knowledge in students' strategy use and reading achievement on the basis of the similar sizes of correlations across countries. In another study, Artelt et al. (2009) tested the criterion validity of metacognitive knowledge with a sample of 15-year-old German students from different school tracks (n = 174). Here, the correlation between the metacognitive knowledge and reading competence was not significant among students from the Gymnasium track (academic track, n = 85). To sum up, there was some support for the criterion validity of the metacognitive knowledge scored with the pair-wise method, but these studies did not explicitly compare this scoring method with a scoring method using raw Likert-scale responses, and evidence on whether these measures are psychometrically comparable across cultural groups is lacking.

Since there is no solid empirical evidence on cross-cultural comparability and validity to guide the scaling of metacognitive knowledge in reading, it remains unclear whether the pair-wise comparison scoring method adopted for PISA 2009 produced more comparable and valid estimates than the traditional Likert-scale scoring method.

Cross-cultural comparability of scale scores

In cross-cultural research, conceptual and psychometric comparability of measures should be demonstrated before any comparative inference is made (Van de Vijver & Leung, 1997, 2000). In large-scale assessments, score differences of a measure across countries may reflect genuine differences in the target construct (i.e., target variance), but also non-target variance due to measurement bias stemming from differences in understanding the construct (i.e., construct bias), differences in sampling, instrument characteristics, and administration procedures (i.e., method bias), and different psychological meaning of item content (i.e., item bias). These measurement biases can jeopardise the comparability of the measure, resulting in measurement non-invariance and preventing valid comparisons. For instance, in their study of students 'self-concept in reading in 48 countries, Authors (2019) reported a lack of full comparability for both perception of competence and perception of difficulty (the two subdimensions of self-concept) and highlighted bias raising from item keying (e.g., positive and negative wording) to differentially affect their

correlations with reading achievement at the country level. In the case of metacognitive knowledge, the pair-wise scoring method as described in the PISA technical report and used by Artelt et al. (2009), Artelt & Schneider (2015)) and the Likert-scale response scoring may respectively increase or reduce different types of measurement bias, thus showing higher or lower cross-cultural comparability.

Metacognition knowledge, control strategy and reading

The three most commonly used criteria for external validity of metacognitive knowledge are reading competence, motivation, and control strategy use. Reading competence has been repeatedly found to be positively related to metacognitive knowledge (Cubukcu, 2008; Taraban et al., 2004). Students' reading performance can be enhanced through learning how to identify and use effective strategies (Pressley et al., 1995). Training students to use metacognitive learning strategies helps them to develop their reading skills and raise their language proficiency levels (Carrell et al., 1998; Green & Oxford, 1995; Palincsar, 1986). As mentioned before, a previous study using the 2009 PISA data also considered reading competence to be the most important criterion for validity (Artelt & Schneider, 2015).

Many studies on self-regulated learning revealed a positive correlation between learning motivation and metacognition/metacognitive strategy use (Pintrich & de Groot, 1990; Roeschl-Heils et al., 2003; Wolters, 1999). Lau and Chan (2003) reported that poor readers applied fewer metacognitive or self-regulated strategies and had lower reading motivation, especially intrinsic motivation. Thus, metacognitive knowledge in reading should be related to motivation.

Furthermore, metacognitive knowledge is correlated with the use of metacognitive strategies (Artelt & Schneider, 2015; Goswami, 2008; Hacker et al., 2009; Schneider & Artelt, 2010). In PISA 2009, different learning strategies were investigated but only the control strategy can be considered to be a metacognitive strategy. The control strategy involves planning, monitoring, and regulation when studying (OECD, 2010). It is essential for effective learning independent of task type and contextual factors (Flavell, 1976; Sitzmann & Ely, 2011). Previous empirical studies using the 2009 PISA data also provided supportive evidence for the associations among control strategy use, reading performance, as well as metacognitive knowledge (Artelt & Schneider, 2015; OECD, 2010).

The present study

Goals of the study

The present study investigates the comparability and validity of metacognitive knowledge based on two scoring methods, i.e. the pair-wise on the one hand and the use of raw Likert-scale scores on the other hand. Specifically, we evaluate the two scoring methods on two criteria: cross-cultural comparability in the measurement and criterion validity when linked with reading competence, motivation, and control strategy use.

In addition, as metacognition knowledge, especially conditional knowledge, refers to the selection of the most adequate strategies according to the tasks, task types might have an impact on the relationship between metacognitive knowledge and the validity criteria (Hakel, 1968). In the above mentioned validation study on metacognitive knowledge in

PISA (Artelt & Schneider, 2015), a composite score of metacognitive knowledge was created, which might have masked task specificity in the associations between metacognitive knowledge and the criterion measures. In our study, the task types are analysed separately to investigate task specificity and add nuance to the existing evidence.

Research questions

In the present study, we compared two scoring methods of metacognitive knowledge. The first method is the original method based on pair-comparisons of original items applied in PISA 2009 (M1) and the other uses the direct Likert rating of selected items (M2). A detailed description of each method is provided in the method section. The two methods are compared on the following criteria: 1) measurement invariance, and 2) criterion validity when they are related to reading competence, motivation in reading, and the use of control strategy. The validity check is conducted separately for the two reading tasks (ST41: remembering and understanding; ST42: summarising). The following research questions are investigated:

(1) Which scoring method for metacognitive knowledge (M1: pair comparison or M2: direct rating) produces data that are more cross-culturally comparable?
(2) Which scoring method for metacognitive knowledge shows higher criterion validity in relation to reading competence, motivation in reading and the use of control strategy?
(3) Does task type influence the cross-cultural comparability and the validity of the metacognitive knowledge measurement with the two scoring methods?

Method

Data

The present study is a secondary analysis of the 2009 PISA student data. PISA in 2009 assessed 15-year-old students' reading, mathematics and science achievement in 34 OECD and 21 partner countries. In this study, we used data from the 34 OECD countries with a total sample size of 298,454 students. Due to differences in the sampling frame and national sample extensions, sample sizes per country varied from 3,646 students in Iceland to 38,250 in Mexico (OECD, 2012).

Measures

Metacognitive knowledge in reading

Metacognitive knowledge was measured with two reading scenarios (short vignettes): remembering and understanding (ST41) and summarising (ST42). Students were asked to rate the usefulness of different strategies on a 6-point Likert scale. Figure 1 presents all items for the scales. In the 2009 PISA international database, derived variables from the two scales were labelled UNDREM (understanding and remembering) and METASUM (summarising).

Reading task: You have to understand and remember the information in a text.

How do you rate the usefulness of the following strategies for understanding and memorising the text?

Possible strategy	Score					
	Not useful at all					*Very useful*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
a) I concentrate on the parts of the text that are easy to understand	☐1	☐2	☐3	☐4	☐5	☐6
b) I quickly read through the text twice	☐1	☐2	☐3	☐4	☐5	☐6
c) After reading the text, I discuss its content with other people	☐1	☐2	☐3	☐4	☐5	☐6
d) I underline important parts of the text.	☐1	☐2	☐3	☐4	☐5	☐6
e) I summarise the text in my own words	☐1	☐2	☐3	☐4	☐5	☐6
f) I read the text aloud to another person	☐1	☐2	☐3	☐4	☐5	☐6

Reading task: You have just read a long and rather difficult two-page text about fluctuations in the water level of a lake in Africa. You have to write a summary.

How do you rate the usefulness of the following strategies for writing a summary of this two-page text?

Possible strategy	Score					
	Not useful at all					*Very useful*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
a) I write a summary. Then I check that each paragraph is covered in the summary, because the content of each paragraph should be included	☐1	☐2	☐3	☐4	☐5	☐6
b) I try to copy out accurately as many sentences as possible	☐1	☐2	☐3	☐4	☐5	☐6
c) Before writing the summary, I read the text as many times as possible	☐1	☐2	☐3	☐4	☐5	☐6
d) I carefully check whether the most important facts in the text are represented in the summary	☐1	☐2	☐3	☐4	☐5	☐6
e) I read through the text, underlining the most important sentences. Then I write them in my own words as a summary	☐1	☐2	☐3	☐4	☐5	☐6

Figure 1. Assessment of metacognitive knowledge using two reading scenarios (ST41: understanding and remembering ST42: summarising) in PISA 2009.

Scoring method 1 (M1) and item selection based on expert rating

This scoring method was applied in the 2009 PISA official reporting. It measures students' metacognitive knowledge based on their relative ratings of pairs of strategies (Artelt & Schneider, 2015; OECD, 2012) with a three-step process. Metacognitive strategic knowledge refers to the ability to judge the usefulness of some strategies in comparison to the others (certain strategies are more useful than other strategies), instead of their absolute usefulness (these strategies are more or less useful).

Step 1: Students were asked to rate the usefulness of several strategies related to reading, as shown in Figure 1. Moreover, 68 reading experts from 42 countries rated all the strategies using the same response format.

Step 2: Multiple pair-wise comparisons can be constructed based on the original items (e.g., an exhaustive pair-wise comparison of six strategies would produce 15 possible pairs, resulting in 30 possible relations). For the 11 original items, only 17 ordered relations were finally agreed upon by at least 80% of the experts and were used as criteria in the coding. Nine pairs from the strategies in the remembering and understanding task, based on the rule that strategies stated in item 3, item 4, and item 5 were preferred to those of item 1, item 2, and item 6. For instance, if a student gives a lower rating for item 1 than item 3 (in accordance with the experts' ratings), a score of 1 is assigned for this pair-wise comparison, otherwise a score of 0 is assigned. Similarly, eight pairs were determined for the summarising task based on the rule: (item 4, item 5) > (item 1, item 3)> item 2. Students' responses were coded accordingly. Through this step the 11 original items (six items for the remembering and understanding task and five items for the summarising task) were recoded into 17 dichotomous items, and theoretically students could score between 0 and17 for the two tasks in total. All the pair relations agreed by experts are listed in Table 1.

Step 3: Finally, students' mean scores were calculated and further standardised into scores with OECD mean of zero and standard deviation of one. A higher score indicates a closer alignment with experts and thus a higher level of metacognitive knowledge. The median reliability (Cronbach's alpha) across countries was.80 and .77 for the two tasks respectively.

Scoring method 2 (M2) and item selection based on expert rating

This scoring method aims at identifying the useful strategies agreed upon by experts and operationalises metacognitive knowledge as the judgement of the usefulness of strategies.

Table 1. Pair relations agreed by experts in the pair-wise comparison method (M1).

Question	ST41 (remembering and understanding)	ST42 (summarising)
Pair relations	ST41c* >ST41a	ST42d>ST42a
	ST41c> ST41b	ST42d > ST42c
	ST41c> ST41f	ST42d > ST42b
	ST41d >ST41a	ST42e >ST42a
	ST41d> ST41b	ST42e > ST42c
	ST41d> ST41f	ST42e > ST42b
	ST41e >ST41a	ST42a >ST42b
	ST41e> ST41b	ST42c >ST42b
	ST41e> ST41f	

Note. *item number corresponding to the strategy number in Figure 1.

Data from the same reading experts as in step 1 of M1 was used. Experts' ratings on specific strategies show both similarities and differences. Only items which received a majority consensus among experts were selected as indicators of metacognitive knowledge in each reading task. These items satisfy the condition that most of the experts (at least 80%) rated these items as 'useful strategies' (with a mode 6 or 5) or 'not useful strategies' (with a mode 1); and the standard deviation (SD) of their ratings was less or equal to 1 (Witner & Tepner, 2011). A summary statistics of experts' rating is presented in Table 2. Based on these criteria, three items of the remembering and understanding task and three items of the summarising task were chosen.

Five of the six items were considered to be useful strategies according to the experts. Four items had a mode 6, one item had a mode 5 (with 29 expert ratings of 5 and 24 expert ratings of 6, thus this item was judged to be quite useful as well) and one item was considered to be not useful as a strategy (with a mode 1, which was later reverse-coded). The mean rating of the three items for the remembering and understanding task and the three items for the summarising task were calculated. In order to be comparable with derived variables in M1, the means were also standardised and transformed to scores with OECD mean of zero and standard deviation of one. Similarly, two new derived variables comparable to UNDREM and METASUM were constructed. They are referred to as UNDREM-D and METASUM-D henceforth. 'D' indicates direct evaluation of the strategies. A higher score here also indicates a judgement closer to experts and thus a higher level of metacognitive knowledge. The median reliability across countries was .70 and .41 for the two tasks respectively.

Reading competence

Reading competence was measured by the reading literacy test, which assessed students' competence in accessing and inferring information, forming a coherent interpretation, and reflecting upon the form and content of authentic reading material (OECD, 2012). The reading achievement score was represented by five plausible values (PV1READ to PV5READ), which were a selection of likely proficiencies randomly drawn from the marginal posterior of the latent distribution for each student. They have a mean value of 500 and standard deviation of 100. With the extensive measure and sophisticated scaling method for the cognitive test in PISA, we assume in this study that reading competences can be validly compared across countries.

Motivation in reading

Motivation in reading was assessed in the student questionnaire. The scale JOYREAD (joy/like reading), consisting of 11 items with 4-point Likert-scale (1 = *strongly disagree*; 4 = *strongly agree*), had high internal consistency values across countries (median reliability = .90). One sample item reads: 'Reading is one of my favourite hobbies'. All items were calibrated and

Table 2. IAgreement of *expert ratings on the selected items in M2.*

	ST41c	ST41d	ST41e	ST42b	ST42d	ST42e
Content	Discuss	Underline	Summarise	Copy	Check	Summarise
Mode	6	5	6	1	6	6
Percentage	95.6%	95.6%	94.1%	100%	100%	91.2%
M (SD)	5.26 (1.03)	5.09 (.84)	5.38 (.91)	1.15 (.40)	5.51 (.70)	5.21 (1.04)

scaled according to an item response theory-based scaling method and the final score was represented by Weighted Likelihood Estimates (WLE), transformed into an international metric with an OECD mean of zero and standard deviation of one (OECD, 2012).

Control strategy use

Learning strategies were assessed on the basis of students' self-reports of strategy use. The control strategy (CSTRAT) consists of five items with response options ranging from 1 (*almost never) to* 4 (*almost always*) (OECD, 2012). An example of CSTRAT item is: 'When I study, I start by figuring out what exactly I need to learn'. This scale has a median reliability of .75 across countries. The scale score was standardised and transformed to have an OECD mean of zero and standard deviation of one (OECD, 2012).

Statistical analysis

To assess the cross-cultural comparability of the two scoring methods, a multi-group confirmatory factor analysis (MGCFA) was conducted to items of each scale. MGCFA is the most frequently applied statistical test for cross-cultural comparability of scales (Cieciuch et al., 2014). Three main levels of invariance can be distinguished and tested applying this approach (Van de Vijver & Leung, 1997): Configural, metric and scalar invariance. Configural invariance indicates that the items cover the facets of the construct adequately in all groups; metric invariance indicates the same factor loadings across groups, which allows for comparisons of within-group associations among variables across groups, but not for the comparison of scale mean scores. Scalar invariance implies that items have the same loadings and intercepts across groups, which allows for comparisons of the scale mean scores across groups. By identifying the invariance level of the scale constructed by different scoring methods, the cross-cultural comparability of these scoring methods can be assessed. Configural, metric and scalar invariance models were tested with the data from all 34 OECD countries with the R package Lavaan (Rosseel, 2012). Model fit of MGCFA was evaluated by the Tucker Lewis Index (TLI) (acceptable above .90), Comparative Fit Index (CFI) (acceptable above .90), and Root Mean Square Error of Approximation (RMSEA) (acceptable below .08) (Cheung & Rensvold, 2002). The acceptance of a more restrictive model is based on the change of CFI and RMSEA values. In the contexts of large-scale assessment with dozens of cultures, Rutkowski and Svetina (2014) proposed to set the cut point of change of CFI to .02 and that of RMSEA to .03 from configural to metric model, and from metric to scalar model the changes of both CFI and RMSEA should be within .01.

Next, correlation analyses of the metacognitive knowledge and the validity measures were conducted within each country. To obtain unbiased estimates from complex large-scale international surveys such as PISA, we carried out the correlational analysis with the IDB analyser (IDB, 2009), which can deal with specific data features of the data set such as sampling and replication weights and the estimation using plausible values. Ahead of the correlation analysis, measurement invariance of the two other self-reported scales (motivation in reading and control strategy use) was also analysed to ensure the comparability of correlations across countries (model fit reported in the result section).

All analyses were conducted separately for the two tasks, thus we can check for task specificity. In addition, to compare sizes of correlations across scoring methods, t-tests were conducted for each country using 'paired.r' function from R package 'psych'. This function can be used when the to-be-compared correlations were dependent on each other.

Results

Measurement invariance tests of metacognitive knowledge using M1 and M2

We tested measurement invariance with the items of the scales based on each scoring method. The model fit results are shown in Table 3. For M1, fit indexes for both tasks showed a poor model fit at the configural level (CFI and TLI lower than .6, RMSEA higher than .2, and SRMR higher than .1) as well as at the metric level (CFI and TLI lower than .5, RMSEA higher than .2, and SRMR higher than .1), indicating a lack of comparability at the configural level with this scoring method. For M2, there were only three items for each derived variable and the configural model was saturated (model fit cannot be evaluated). However, the metric invariance model fitted relatively well (CFI and TLI higher than .95, RMSEA and SRMR lower than .08), indicating that across countries all items were related to the construct in a similar manner. The fit of the scalar invariance model for both tasks was significantly poorer than the metric invariance model, thus could not be accepted. Taken together, we can conclude that UNDREM-D and METASUM-D reached metric invariance across countries whereas UNDREM and METASUM did not achieve configural invariance. Therefore, there is clear evidence that M2 showed higher cross-cultural comparability than M1.

Correlations with reading competence

The zero-order correlations between metacognitive knowledge and all three validity measures in the two reading tasks as produced in the IDB analyser are presented in Table 4.

Table 3. Model fit of the multigroup confirmatory factor analyses.

	CFI	TLI	RMSEA	SRMR	ΔCFI	Δ RMSEA
Understanding and remembering Task-M1						
Configural	.479	.305	.274	.131		
Metric	.475	.456	.242	.133	−.004	−.032
Scalar	.452	.536	.224	.138	−.023	−.018
Summarising Task-M1						
Configural	.509	.313	.290	.144		
Metric	.504	.481	.252	.147	−.005	−.038
Scalar	.481	.567	.230	.152	−.023	−.022
Understanding and remembering Task-M2						
Metric	.990	.984	.054	.023		
Scalar	.894	.918	.124	.066	−.096	.043
Summarising Task-M2						
Metric	.977	.964	.068	.034		
Scalar	.862	.894	.117	.073	−.015	.049

Note. The model fit was evaluated by the Tucker Lewis Index (TLI) (acceptable above.90), Comparative Fit Index (CFI) (acceptable above.90), and Root Mean Square Error of Approximation (RMSEA) (acceptable below.08).

Table 4. Correlations of metacognitive knowledge with reading competence, motivation in reading and control strategy use in the remembering task and Summarising task.

	r Metacognitive knowledge in reading and **remembering** task (ST41) *with*						*r* Metacognitive knowledge in reading and **summarising** task (ST42) *with*					
	Reading competence		Motivation in reading		Control		Reading competence		Motivation in reading		Control	
Country	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
Australia	0.40	0.35	0.28	0.36	0.32	0.48	0.47	0.50	0.33	0.41	0.30	0.45
Austria	0.43	0.34	0.26	0.31	0.19	0.34	0.48	0.49	0.28	0.34	0.16	0.28
Belgium	0.48	0.35	0.29	0.30	0.28	0.37	0.53	0.54	0.29	0.35	0.26	0.37
Canada	0.31	0.24	0.21	0.28	0.24	0.38	0.40	0.42	0.26	0.34	0.26	0.38
Chile	0.41	0.35	0.18	0.32	0.21	0.38	0.41	0.47	0.17	0.29	0.13	0.30
Czech	0.41	0.36	0.29	0.33	0.25	0.37	0.52	0.53	0.33	0.38	0.24	0.33
Denmark	0.43	0.36	0.29	0.34	0.23	0.38	0.44	0.47	0.30	0.37	0.22	0.33
Estonia	0.39	0.28	0.23	0.24	0.16	0.27	0.43	0.44	0.26	0.33	0.12	0.24
Finland	0.42	0.39	0.36	0.43	0.28	0.37	0.49	0.54	0.37	0.47	0.27	0.38
France	0.40	0.33	0.24	0.31	0.26	0.40	0.47	0.52	0.26	0.34	0.27	0.39
Germany	0.46	0.33	0.26	0.26	0.23	0.34	0.50	0.54	0.29	0.35	0.18	0.31
Greece	0.21	0.16	0.19	0.29	0.17	0.35	0.35	0.43	0.20	0.35	0.16	0.33
Hungary	0.39	0.25	0.25	0.30	0.15	0.33	0.51	0.51	0.30	0.39	0.11	0.27
Iceland	0.35	0.30	0.26	0.27	0.23	0.37	0.44	0.48	0.28	0.33	0.20	0.33
Ireland	0.36	0.25	0.22	0.23	0.24	0.35	0.41	0.43	0.24	0.33	0.21	0.32
Israel	0.35	0.26	0.12	0.24	0.13	0.37	0.43	0.46	0.19	0.29	0.10	0.31
Italy	0.38	0.25	0.24	0.26	0.22	0.38	0.44	0.46	0.23	0.31	0.20	0.35
Japan	0.38	0.37	0.17	0.23	0.21	0.32	0.51	0.56	0.23	0.30	0.23	0.33
Korea	0.42	0.48	0.21	0.34	0.25	0.42	0.51	0.55	0.22	0.31	0.30	0.40
Luxembourg	0.41	0.27	0.27	0.31	0.29	0.40	0.47	0.49	0.30	0.40	0.25	0.38
Mexico	0.34	0.23	0.15	0.29	0.16	0.34	0.42	0.42	0.15	0.26	0.16	0.29
Netherlands	0.45	0.34	0.26	0.32	0.29	0.36	0.50	0.53	0.26	0.32	0.20	0.31
New Zealand	0.38	0.31	0.24	0.36	0.29	0.44	0.48	0.50	0.28	0.38	0.27	0.41
Norway	0.38	0.35	0.25	0.35	0.26	0.40	0.44	0.47	0.25	0.35	0.26	0.36
Poland	0.32	0.31	0.22	0.30	0.17	0.36	0.46	0.49	0.27	0.34	0.19	0.32
Portugal	0.44	0.39	0.24	0.36	0.29	0.45	0.51	0.57	0.26	0.38	0.29	0.45
Slovak	0.35	0.30	0.23	0.28	0.25	0.38	0.47	0.47	0.23	0.28	0.22	0.31
Slovenia	0.41	0.35	0.27	0.34	0.22	0.41	0.46	0.50	0.27	0.37	0.16	0.32
Spain	0.32	0.29	0.22	0.29	0.23	0.39	0.43	0.47	0.26	0.32	0.24	0.38
Sweden	0.43	0.38	0.32	0.39	0.27	0.42	0.46	0.50	0.30	0.39	0.20	0.34
Switzerland	0.49	0.37	0.30	0.35	0.27	0.41	0.51	0.53	0.31	0.38	0.23	0.35
Turkey	0.31	0.15	0.14	0.31	0.13	0.38	0.38	0.38	0.15	0.29	0.10	0.29
UK	0.35	0.25	0.22	0.27	0.24	0.39	0.42	0.44	0.25	0.31	0.18	0.33
US	0.34	0.22	0.25	0.30	0.26	0.40	0.39	0.42	0.24	0.33	0.22	0.37
Median	0.39	0.32	0.24	0.30	0.24	0.38	0.46	0.49	0.26	0.34	0.22	0.33
Mean	0.39	0.31	0.24	0.31	0.23	0.38	0.46	0.49	0.26	0.34	0.21	0.34
SD	0.06	0.07	0.05	0.04	0.05	0.04	0.04	0.05	0.05	0.04	0.06	0.05

Note. All the correlations were significant at the.05 level. All the correlations between metacognitive knowledge and plausible values were significant at.01 level.

Given the large sample size, the correlation coefficients listed in Table 4 were all significant at alpha = .05. Both M1 and M2 metacognitive knowledge correlated positively with reading competence, pointing to good validity. In the remembering task, metacognitive knowledge scored with M1 correlated more strongly with the reading competence than these scored with M2 ($2.79 < t < 26.53$, $p < .05$) in all countries except for Japan, Poland and Korea. Korea was the only country in which M2 showed a higher correlation and reading competence than M1 ($t = -5.72$, $p < .001$). In both Japan and Poland, there was no significant difference between M1 and M2. On the contrary, in the summarising task, for 26 out of 34 countries metacognitive knowledge scored with M2 correlated more strongly with reading competence than M1 ($-10.23 < t < -2.10$, $p < .05$). For Hungary, Mexico, Slovak Republic and Turkey, no difference was found between M1 and M2.

Correlations with reading motivation

In order to get a valid correlation between metacognitive knowledge measurement and the criteria measurements, we also tested the cross-cultural comparability of the scale criterion measures. For the reading motivation scale, a MGCFA supported metric invariance: CFI = .90, TLI = .90, RMSEA = .10, SRMR = 0.072, the drop of CFI value from the configural model was only .01, and RMSEA was only reduced by .004 from the configural to the metric invariance model.

As shown in Table 4, the correlations between reading motivation and metacognitive knowledge were positive, and these using M2 were generally higher than scores using M1. Further comparisons of correlations showed that in 29 out of 34 countries, the differences were significant ($-26.24 < t < -3.21$, $p < .05$) for the remembering and understanding task. For the other five countries (namely Belgium, Estonia, Iceland, Ireland and Germany), the differences were not significant, while no country showed an opposite pattern. In the summarising task, the correlations with M2 were higher than those with M1 and all differences were significant ($-29.63 < t < -4.73$, $p < .05$).

Correlations with control strategy use

For the control strategy use scale, the MGCFA showed acceptable metric invariance: CFI = .98, TLI = .98, RMSEA = .047, SRMR = 0.029; change of CFI from configural to metric model was .006, and that of RMSEA was.003. This supports the valid comparison of correlations across cultures.

All the correlations between metacognitive knowledge and the control strategy were positive and significant ($p < .05$). Control strategy use correlated more strongly with metacognitive knowledge using M2 than that using M1 (for all the countries in the remembering and understanding task: $-34.29 < t < -7.01$, $p < .05$; and for all the countries in the summarising task: $-39.42 < t < -10.16$, $p < .05$).

Task specificity

The correlational results reported above revealed that scores of metacognitive knowledge from both scoring methods for both tasks had positive correlations with reading competence, motivation, and the control strategy. However, the strength of correlations tended to differ across the scoring methods and across the tasks, especially for the correlation with reading competence. As shown before, for the remembering and understanding task, M1 scores correlated more strongly with reading competence than M2 scores ($2.79 < t < 26.53$, $p < .05$); whereas for the summarising task the opposite pattern emerged: the correlations using M2 were generally higher than those using M1 ($-10.23 < t < -2.10$, $p < .05$). Besides the general pattern, it should be noted that Korea showed a different correlation pattern in comparison with other countries, i.e. a higher correlation of M2 than that of M1 for the remembering task.

Discussion

We set out to investigate the measurement of metacognitive knowledge with two different scoring methods (M1, a pair-wise comparison of strategies and M2, scaling with raw Likert ratings on selected strategies) in different reading tasks with data from 34 OECD countries. Our main findings were (1) M2 showed higher cross-cultural comparability than M1 (RQ1); (2) metacognitive knowledge scored with the two scoring methods showed different criterion validity (RQ2) and some task specificity was revealed (RQ3). Specifically, with reading competence, M1 showed stronger correlations in most countries in the understanding and memorising task than M2, whereas M2 showed stronger correlations in the summarising task than M1. M2 in contrast to M1 tended to show stronger correlations with motivation in reading, and with the control strategy in both tasks across countries. Given these findings, we can conclude that in terms of measurement comparability across countries, M2 outperformed M1. However, the criterion validity of metacognitive knowledge is more nuanced than expected. We discuss each of the findings and their implications.

Measurement comparability

First, rating responses (M2) tended to show better psychometric properties than scores derived from pair-wise comparisons. Although response formats other than direct Likert-scale ratings may reduce response style bias and enhance cross-cultural comparison (Authors, 2013, 2015), this advantage of cross-cultural invariance expected from M1 was not supported by our measurement invariance tests: M2 reached metric invariance while M1 showed poor fit for all the invariance models. As cross-cultural comparability is an important criterion for comparative studies such as PISA, M2 showed better measurement quality. According to previous literature, the evidence of the generalisability of M1 (Artelt & Schneider, 2015) was based on similarities of correlations of metacognitive and reading competence across countries without first conducting invariance tests. Our study provides statistical testing evidence leading to caution the generalisability of M1, given that the rescored data based on pair comparisons had rather poor cross-cultural comparability. We advocate the call to always empirically check the measurement invariance of target scales before using the scale scores for further substantive cross-cultural analysis (e.g., Authors., 2019; Vieluf et al., 2013).

Differential criterion validity

Reading achievement, motivation in reading and the control strategy use are important correlates of metacognitive knowledge (Cubukcu, 2008; Lau & Chan, 2003; Sitzmann & Ely, 2011; Taraban et al., 2004). We used these variables as criteria to compare the validity of metacognitive knowledge scored with M1 and M2. Artelt and Schneider (2015) reported medium to high correlations between metacognitive knowledge scored by M1 and reading competence and control strategy use. Our results were in general in line with their findings (using both M1 and M2). However, two differential patterns are noteworthy when comparing M1 and M2 in their relation to the three validity measures in the two tasks.

First, reading competence had in general higher correlations with metacognitive knowledge scored in M1 than M2 in the understanding and remembering task, whereas the reverse was true for the summarising task. Hakel (1968) argued that certain characteristics of tasks such as complexity and cognitive load influence the measurement. The two reading tasks differ in these aspects. Specifically, the understanding and remembering task is less complex than the summarising task, and it is relatively easier to achieve (whereas summarising requires first to understand and then to exert extra cognition to elicit key information based on this understanding) (Artelt et al., 2009). In the understanding and remembering task, each item is straightforwardly stated and only consists of one specific strategy (e.g., 'I underline important parts of the text'), and there seems to be a clear differentiation between more useful and less useful strategies (i.e., underlining/summarising vs. reading). Therefore, M1 based on this pair comparisons elicited more useful information than the M2 raw Likert ratings of the three chosen items in the understanding and remembering task, and captured more shared variance with reading achievement than in M2. In the summarising task, listed strategies are more complex. Some items even include multiple steps, i.e., 'I write a summary, then I check that each paragraph is covered in the summary, because the content of each paragraph should be included', which makes it difficult to discriminate one set of strategies against another. Therefore, M2 Likert ratings of selected strategies in this task captured more shared variance with reading achievement than M1.

Despite a high consensus among experts for all the selected items in M2, the ratings on these strategies from students with the highest reading performance scores were not necessarily aligned with expert ratings. For example, in the remembering and understanding task, students in OECD countries with reading competence scores higher than 625.61 (at the highest two levels of proficiency) rated the underlining strategy (item d) as the most effective strategy (a mode rating of 6) while the discussion strategy (item c) was rated as somewhat less effective (a mode rating of 5). In contrast, the experts considered discussion (a mode rating of 6) to be more effective than underlining (a mode rating of 5). Country-level analysis showed that the discrepancy between high achieving students and experts can be observed for most of the countries in the OECD (as opposed to countries with most high achievers). The mismatch in ratings between experts and students with higher performance scores might explain why M2 for the understanding and remembering task had a lower correlation with reading achievement. Given that metacognitive knowledge includes knowledge about the task, strategies and learners themselves (Artelt et al., 2009; Flavell, 1976), it is reasonable that the judgement on useful strategies is contingent on the characteristics of the tasks and learners' experiences with the strategies.

Secondly, in relation to motivation and the control strategy, M2 consistently showed higher criterion validity than M1 in both tasks. In previous studies, metacognitive knowledge in reading was usually measured by investigating participants' direct ratings on their actual strategy use, either directly after their reading or of their daily reading (Lau & Chan, 2003; Schraw & Dennison, 1994; Van Gelderen et al., 2004). For the measurement of metacognitive strategies in reading, questionnaires using Likert-scales were used (Pintrich et al., 1991; Weinstein et al., 1987). M2 is based on the same method (Likert-scale rating). A methodological caveat for the higher correlations between scales using Likert responses is the so-called common method bias (Podsakoff et al., 2003). That is, the common response format of the instrument can inflate the correlations between

scale scores, and it is difficult to distinguish the overlap of variance due to substance or due to the common method. With Likert scales, the presence of response styles (i.e., the systematic tendency to respond on the basis other than the target construct) is a well-known phenomenon, and respondents may exhibit stable tendency to endorse certain categories in items of different constructs, which can 'inflate' the correlations. It would be interesting to control for the common method variance in Likert-scale data and elicit correlations based on substance.

The lower correlation of M1 with motivation and control strategy use might also be due to its indirect measurement and lack of construct validity (configural invariance not achieved). M1 was based on the 'relational strategy knowledge' (Artelt & Schneider, 2015), which was supposedly an innovative measurement with great potential to enhance the validity. Such measurement would benefit from a testing situation in which the participants are explicitly asked to compare several strategies and choose the preferred strategies for a specific reading task, as was applied in previous scenario-based instruments (e.g., Metacomprehension Strategy Index by Schmitt, 1990). This is not how these items were administered in PISA (experts and students were only asked to rate the usefulness of several strategies one after another). Factors such as positioning of the strategies or students' awareness of strategy comparison and selection can influence their ratings. Students' indirect ranking of these reading strategies, as captured in M1, does not necessarily reflect their opinion about the relative ranking of the strategies and further measure their levels of metacognitive knowledge accurately.

Apart from the method and task specificity, country heterogeneity and the interaction between task type and country were observed. The most peculiar finding was that Korea showed a different pattern for the correlations between metacognitive knowledge and reading competence in remembering task (ST41). Artelt and Schneider (2015) also highlighted Korea as having a divergent correlational pattern: the correlations between reading performance and elaboration strategy, as well as between reading performance and memorisation strategy use, were positive and the highest for Korea in their analysis. These results point to a possible influence of cultural traditions which might also be related to different practices in education (Kember, 1996; Marambe et al., 2012; Marton et al., 2005): Eastern Asian students with high degrees of metacognitive knowledge might perceive memorisation or elaboration to be appropriate strategies especially for remembering and understanding tasks. The memorisationstrategy might require careful reading, comprehension and interpretation for Asian students (Baumgart & Halse, 1999). Still, it would be interesting to see if this pattern can be replicated for other Asian, Confucian-based countries.

Limitations and further directions

There are a few limitations to our study. First, the strategies listed under each task did not cover the breadth of all metacognitive strategies, thus the more generic metacognitive knowledge is elusive. A comprehensive list of strategies in different types of tasks may reveal more valid cultural and individual variations in metacognitive knowledge in reading. Secondly, the common scale applied by M2 and the other criteria could inflate the correlation among them. A closer inspection of the influence of common method variance is needed in the future. Furthermore, our data were from students' self-reports

in a cross-sectional design. The validity and effectiveness of metacognitive knowledge in reading would benefit from data of multiple sources and research designs. Moreover, cross-cultural difference in metacognitive knowledge was not our focus, but country–specificity especially from a non-Western point of view (e.g., Nardi, 2008) is worth investigating if we were to find a more accurate measurement for metacognitive knowledge in reading for different countries. In line with findings on the difference between experts and high-achieving students, it is also worth investigating qualitatively the perception of certain strategies (e.g., memorisation, peer-discussion.) among respondents of various groups (e.g., high or low achievers, teachers, educational researchers) in order to develop a framework for metacognitive knowledge in reading for more comparable and valid cross-cultural studies.

Conclusion

Our study contributes to understanding the complexity of the measurement of metacognitive knowledge in cross-cultural studies. Drawing on the comparison of two scoring methods, we draw two main conclusions. First, metacognitive knowledge scored with the raw Likert-scale responses, in comparison to the pair-wise comparison method, had higher comparability (i.e., metric equivalence), which points to its advantage in ensuring valid comparisons of within-culture associations across cultural groups, and the necessity to always check the measurement comparability in cross-cultural research. Secondly, findings on the differential criterion validity from the two scoring methods presented a more nuanced picture: results might be related to the complexity of specific strategies (possibility of reaching high consensus for pair-wise comparison), discrepancy between ratings from experts and those from students with higher performance scores, potential confounding of common method bias, and cultural differences in values and educational tradition.

Our findings are not only meaningful for researchers analysing metacognitive knowledge with PISA data and with different scoring methods. They also have implications for policies and practices of educational assessment. Firstly, it has often been argued that psychometric properties including measurement invariance in cross-cultural assessment should be demonstrated (Boer et al., 2018). This may require extensive pre-testing and test adaptation for local contexts for large-scale assessment, and with the main study data, these psychometric property statistics should be reported and comparisons of cultural groups should only be done with invariant data. Awareness of these issues should reach policy-makers, survey implementation teams, analysts and the general public alike. Secondly, when multiple scoring methods are possible, they should be empirically compared and contrasted (as shown in our study). Different scoring methods may be preferred given the research aim and design. For instance, in our study we recommend M2 for higher cross-cultural comparability but M1 outperforms M2 for higher criterion validity when associating metacognitive knowledge with reading achievement in certain tasks (low complexity and clear-cut strategies). There is no one silver bullet that solves all psychometric problems and shows good validity. We also expect that convergence of multiple sources and scoring methods can contribute to the robustness or the nuanced understanding of findings. Particularly for the assessment of metacognitive knowledge in reading, researchers and practitioners who intend to improve the measurement of

metacognitive knowledge in reading in large scale assessment should be aware of: 1) metacognitive knowledge involves many aspects and multiple steps and should be defined clearly and tested accordingly;. 2) factors such as task type, criterion group, scale type, and culture might influence the definition/operationalisation and the validity of the measurement. To maximise the ecological validity (i.e., balance comparability and criterion validity) of metacognitive knowledge, the above mentioned aspects should be taken into consideration.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Ji Zhou has studied in Beijing Normal University and obtained her PhD in educational psychology in Ludwig-Maximilians-University München, Germany. Her research interest lies in cross-cultural assessment, learning motivation, metacognition, and the methods for large scale analysis.

Jia He obtained her PhD in cross-cultural psychology in Tilburg University, the Netherlands. Her research includes cross-cultural assessment and advanced psychometric methods. She is interested in modern research methods such as item response theory-based scaling, structural equation modelling, multilevel analysis, and Bayesian statistics.

Dominique Lafontaine is a Full Professor of Educational Science at the University of Liège and is director of the research centre 'Analysis of systems and practices in education' (aSPe). D. Lafontaine has a strong background in the field of reading literacy, comparative studies, teaching and learning processes, development of cognitive and non-cognitive instruments, effectiveness and equity of education systems. D. Lafontaine has been a member of the PISA-Reading expert group since 1998 and was a member of the Questionnaire expert group for PISA 2018.

ORCID

Ji Zhou http://orcid.org/0000-0001-8318-1457
Jia He http://orcid.org/0000-0001-7310-4861
Dominique Lafontaine http://orcid.org/0000-0003-1497-4634

References

Akturk, A. O., & Sahin, I. (2011). Literature review on metacognition and its measurement. *Procedia - Social and Behavioral Sciences*, *15*, 3731–3736.

Allen, B. A., & Armour-Thomas, E. (1993). Construct validation of metacognition. *The Journal of Psychology*, *127*(2), 203–211. https://doi.org/10.1080/00223980.1993.9915555

Artelt, C., Beinicke, A., Schlagmüller, M., & Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *41* (2), 96–103. https://doi.org/10.1026/0049-8637.41.2.96

Artelt, C., & Schneider, W. (2015). Cross-country generalizability of the role of metacognitive knowledge in students' strategy use and reading competence. *Teachers College Record*, *117*(1), 1–32. http://www.tcrecord.org/Content.asp?ContentId=17695

Authors. (2013).

Authors. (2015).

Authors. (2019).

Baumgart, N., & Halse, C. (1999). Approaches to learning across cultures: The role of assessment. *Assessment in Education: Principles, Policy & Practice*, *6*(3), 321–339. https://doi.org/10.1080/09695949992775

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Carrell, P. L., Gajdusek, L., & Wise, T. (1998). Metacognition and EFL/ESL reading. *Instructional Science*, *26*(1/2), 97–112. https://doi.org/10.1023/A:1003092114195

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology*, *5,* 982. https://doi.org/10.3389/fpsyg.2014.00982

Cubukcu, F. (2008). Enhancing vocabulary development and reading comprehension through metacognitive strategies. *Issues in Educational Research*, *18*(1), 1–11. http://www.iier.org.au/iier18/cubukcu.html

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

Goswami, U. (2008). *Cognitive development: The learning brain*. Psychology Press.

Green, J. M., & Oxford, R. L. (1995). A closer look at learning strategies: L2 Proficiency and gender. *TESOL Quarterly*, *29*(2), 261–297. https://doi.org/10.2307/3587625

Hacker, D. J., Dunlosky, J., & Graesser, A. (Eds.). (2009). *Handbook of metacognition in education*. Taylor & Francis.

Hakel, M. D. (1968). Task difficulty and personality test validity. *Psychological Reports*, *22*(2), 502. https://doi.org/10.2466/pr0.1968.22.2.502

IDB. (2009). *IDB Analyzer (Version 3.1)*. International Association for the Evaluation of Educational Achievement.

Jacobs, J., & Paris, S. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist*, *22*(3–4), 255–278. https://doi.org/10.1080/00461520.1987.9653052

Kember, D. (1996). The intention to both memorise and understand: Another approach to learning. *Higher Education*, *31*(3), 254–341. https://doi.org/10.1007/BF00128436

Lau, K., & Chan, D. W. (2003). Reading strategy use and motivation among Chinese good and poor readers in Hong Kong. *Journal of Research in Reading*, *26*(2), 177–190. https://doi.org/10.1111/1467-9817.00195

Love, S., Kannis-Dymand, L., & Lovell, G. P. (2019). Development and validation of the metacognitive processes during performances questionnaire. *Psychology of Sport and Exercise*, *41*, 91–98. https://doi.org/10.1016/j.psychsport.2018.12.004

Mak, S., Cheung, K., Soh, K., Sit, P., & Leong, M. (2017). An examination of student- and across-level mediation mechanisms accounting for gender differences in reading performance: A multilevel analysis of reading engagement. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *37*(10), 1206–1211. https://doi.org/10.1080/01443410.2016.1242712

Marambe, K., Vermunt, J., & Boshuizen, H. (2012). A cross-cultural comparison of student learning patterns in higher education. *Higher Education*, *64*(3), 299–316. https://doi.org/10.1007/s10734-011-9494-z

Marton, F., Wen, Q., & Wong, K. C. (2005). "Read a hundred times and the meaning will appear … " Changes in Chinese University students' views of the temporal structure of learning. *Higher Education*, *49*(3), 291–318. https://doi.org/10.1007/s10734-004-6667-z

Nardi, E. (2008). Cultural biases: A non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 259–266.

OECD. (2010). *PISA 2009 results: Learning to learn: Student engagement, strategies and practices (Volume III)*.

OECD. (2012). *PISA 2009 technical report*.
Palincsar, A. (1986). Metacognitive strategy instruction. *Exceptional Children*, *53*(2), 118–124. https://doi.org/10.1177/001440298605300203
Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, *8*(3), 293–316. https://doi.org/10.1016/0361-476X(83)90018-8
Pintrich, P., Smith, D., García, T., & McKeachie, W. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. University of Michigan.
Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40. https://doi.org/10.1037/0022-0663.82.1.33
Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879
Pressley, M., Brown, R., El-Dinary, P. B., & Afflerbach, P. (1995). The comprehension instruction that students need: Instruction fostering constructively responsive reading. *Learning Disabilities Research and Practice*, *10*(4), 215–224.
Roeschl-Heils, A., Schneider, W., & van Kraayenoord, C. (2003). Reading, metacognition and motivation: A follow-up study of German students in Grades 7 and 8. *European Journal of Psychology of Education*, *18*(1), 75–86. https://doi.org/10.1007/BF03173605
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257
Säälik, Ü. (2015). Reading performance, learning strategies, gender and school language as related issues – PISA 2009 findings in Finland and Estonia. *International Journal of Teaching and Education*, *3*(2), 17–30. https://doi.org/10.20472/TE.2015.3.2.002
Schmitt, M. C. (1990). A questionnaire to measure children's awareness of strategic reading processes. *The Reading Teacher*, *43*(7), 454–461. https://www.jstor.org/stable/20200439
Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM - the International Journal on Mathematics Education*, *42*(2), 16–149. https://doi.org/10.1007/s11858-010-0240-2
Schraw, G., & Dennison, R. S. (1994). Assessing Metacognitive Awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033
Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, *137*(3), 421–442. https://doi.org/10.1037/a0022777
Taraban, R., Kerr, M., & Rynearson, K. (2004). Analytic and pragmatic factors in college students' metacognitive reading strategies. *Reading Psychology*, *25*(2), 67–82. https://doi.org/10.1080/02702710490435547
Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Sage Publications.
Van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-cultural Psychology*, *31*(1), 33–51. https://doi.org/10.1177/0022022100031001004
Van Gelderen, A., Schoonen, R., de Glopper, K., Hulstjn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first-and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, *96*(1), 19–30. https://doi.org/10.1037/0022-0663.96.1.19
Vieluf, S., Kunter, M., & van de Vijver, F. J. R. (2013). Teacher self-efficacy in cross-national perspective. *Teaching and Teacher Education*, *35*, 92–103. https://doi.org/10.1016/j.tate.2013.05.006
Weinstein, C. E., Palmer, D., & Schulte, A. C. (1987). *Learning and Study Strategies Inventory (LASSI)*. H & H Publishing.

Witner, S., & Tepner, O. (2011). Entwicklung geschlossener Testaufgaben zur Erhebung des fachdidaktischen Wissens von Chemielehrkräften [Development of closed-ended questions to measure the knowledge of chemistry teachers]. *Chimica et ceterae artes rerum naturae didacticae*, *37*, 113–137.

Wolters, C. A. (1999). The relationship between high school students' motivational regulation and their use of learning strategies, effort, and classroom performance. *Learning and Individual Differences*, *11*(3), 281–301. https://doi.org/10.1016/S1041-6080(99)80004-1