Education and Science

Vol 46 (2021) No 205 221-238

Metacognitive Monitoring and Mathematical Abilities: Cognitive Diagnostic Model and Signal Detection Theory Approach *

Oğuz Tahsin Başokçu [1], Mehmet Akif Güzel [2]

Abstract

Besides various in-class assessments, there exist some standardized assessment tools that are administered in several countries, such as PISA (Programme for International Student Assessment) and TIMMS (Trends in International Mathematics and Science Study). The questions' contents, type of responding, grading, and the analyses in these large-scale tests have been diversified in years. In this study, it was aimed to identify the abilities that are measured at PISA mathematics test in a single testing procedure and by utilizing the methods of analyses of Cognitive Diagnostic Model (CDM) as well as Signal Detection Theory (SDT), which have not been used so far in the assessment of these abilities. Therefore, a randomly selected sample of 6th-grade students (*N*=230) in Izmir was tested with a PISA-equivalent 12-item mathematics test, where the items are graded dichotomously (correct vs. incorrect). CDM estimates were calculated by using the Deterministic Input Noisy Output and Gate (DINA) Model. The participants were asked to report whether they thought they could solve the question correctly, guess even if they thought they could not solve the question, and then, rate their confidence levels on the correctness of their answers in turn so as to allow us to measure their "metacognitive monitoring performance" with the SDT method, which refers to the ability to differentiate correct and incorrect responses. In short, a better metacognitive monitoring performance was obtained by measuring how well once could differentiate their correct and incorrect responses with the observation of they prefer reporting and then giving high confidence levels to the actually correct responses and prefer passing to give an answer yet rate lower confidence levels to the actually incorrect responses given as pure guesses. The results showed that CDM fits well to the assessment of PISA test and those who were better at the ability of "reasoning and developing strategies" in particular among four possible abilities detected with CDM ("representing and communicating", "mathematization", "reasoning and developing strategies", "using symbolic and technical language") had also

Keywords

Cognitive Diagnostic Model
Signal Detection Theory
Metacognition
Metacognitive Monitoring
Mathematics
PISA test

Article Info

Received: 06.17.2018
Accepted: 06.02.2020
Online Published: 11.12.2020

DOI: 10.15390/EB.2020.7991

* This article is orally presented at the 17th EARLI (The European Association for Research on Learning and Instruction) Conference.

[1] Ege University, Faculty of Education, Department of Educational Sciences, Turkey, oguzbasokcu@gmail.com
[2] Abdullah Gul University, Faculty of Humanities and Social Sciences, Dept. of Psychology, Turkey, akif.guzel@agu.edu.tr

better metacognitive monitoring performance. The present study, therefore, contributes to the research that investigates what features the ability of better differentiating correct and incorrect responses are actually linked. Based on the results, it is suggested that a better metacognitive monitoring ability is linked to having a better ability of "reasoning and developing strategies" in particular. Additionally, it is suggested that measuring metacognitive monitoring performance at PISA -or even any other possible tests- with the SDT calculation method, that has a relatively straightforward testing procedure, may yield various estimates for the students' abilities measured at the test as well as their related higher-order abilities.

Introduction

The primary objective of the assessment tools of academic performance is to evaluate the level of learners' performance as accurately and precisely as possible. For this purpose, numerous measurement strategies and approaches are used in the educational assessment to detect, for instance, cognitive abilities and academic success (e.g., Bean & Peterson, 1998; Wragg, 2001; Lindblom-Ylanne, Pihlajamaki, & Kotkas, 2006). Various standard tests have also been used in the performance assessment, which is a critical element in education, along with some largely used in-class assessments. For instance, some standard tests are used in many countries to identify how the students' performance in mathematics, science, and language differs between countries and to guide how the educational policies should be directed. The best examples for such tests can be PISA, which was administered first in 2000 and have been under development in terms of its content and grading since then, as well as TIMSS.

In this research, which focuses specifically on the mathematical abilities in the PISA test, it was aimed at determining the level of proficiency shown by the 6th-grade students in the PISA mathematics test by simultaneously using two methods together that have not been used in the assessment of PISA test. The first method is of Cognitive Diagnostic Model and the second one is the method based on Type-2 Signal Detection Theory. The present study targeted to identify the students' mathematical abilities with the CDM's method and to reveal the relationship between students' cognitive abilities measured at the test and their metacognitive monitoring performance, referring to one's ability to differentiate correct and incorrect responses (e.g., Higham, 2002; Higham & Gerrard, 2005; Güzel & Higham, 2013; see also, Karakelle & Saraç, 2010)[1]. In short, the objective was to determine which latent classes that are expected to be defined for the mathematics ability measured at PISA test normatively are related to the metacognitive monitoring ability in a singly testing procedure and with the calculation methods of CDM and Type-2 SDT. The present study is expected to yield a unique contribution to the literature since the existing literature seems to be lacked in terms of not revealing direct observations regarding how well the students are good at recognizing their responses' correctness at PISA tests, which is obtained with Type-2 SDT calculations.

The following sections detail the abilities measured at PISA mathematics test and the calculations methods of CDM and Type-2 SDT respectively.

[1] Since the current study follows the basic assumptions of the Type-2 SDT and so uses its specific calculation methods, it prefers using the term, "metacognitive monitoring" (see also, Higham, 2011) that is used in this specific literature in a standard way instead of using other available terms such as "metacognitive calibration" or "metacognitive accuracy" (see also, Pieschel, 2009).

Mathematical Abilities at PISA

There exist several approaches regarding the "mathematics literacy" (OECD, 2003), which refers to one's ability to identify and comprehend at mathematics and to make well-grounded decisions comparable to their ages as to how mathematics plays a role in their current and future daily, social, and work lives (e.g., Stacey & Turner, 2014). However, PISA and TIMMS investigations use more comprehensive and assessment-based definitions by using the question items that cover possibly all main abilities of mathematics (Albacete et al., 2016; Gierl, Alves, & Majeau, 2010). The PISA test, for instance, which is the main investigation topic of the current study, is an assessment system that is administered once in every three years since 2000 among OECD (The Organization for Economic Co-operation and Development) countries to assess the mathematics and science literacies and the language abilities of the students who are 15-years old (OECD, 2010). These capabilities subsumed under seven categories are defined as are displayed in Table 1 (OECD, 2019).

Table 1. The Capabilities Measured at PISA Mathematics Test

Communicating	– Read, decode, and make sense of statements, questions, tasks, objects, or images, to form a mental model of the situation. – Articulate a solution, show the work involved in reaching a solution, and/or summarize and present intermediate mathematical results. – Construct and communicate explanations and arguments in the context of the problem.
Mathematising:	– Identify the underlying mathematical variables and structures in the real-world problem, and make assumptions so that they can be used. – Use an understanding of the context to guide or expedite the mathematical solving process, e.g. working to a context-appropriate level of accuracy. – Understand the extent and limits of a mathematical solution that are a consequence of the mathematical model employed.
Representation:	– Create a mathematical representation of real-world information. – Make sense of, relate, and use a variety of representations when interacting with a problem. – Interpret mathematical outcomes in a variety of formats in relation to a situation or use; compare or evaluate two or more representations in relation to a situation.
Reasoning and argument:	– Explain, defend, or provide a justification for the identified or devised representation of a real-world situation. – Explain, defend, or provide a justification result processes and procedures used to determine a mathematical result or solution. – Connect the pieces of information to arrive at a mathematical solution, make generalizations or create a multi-step argument – Reflect on mathematical solutions and create explanations and arguments that support, refute, or qualify a mathematical solution to a contextualized problem.

Table 1. Continued

Devising strategies for solving problems	– Select or devise a plan or strategy to mathematically reframe contextualized problems – Activate effective and sustained control mechanisms across a multi-step procedure leading to a mathematical solution, conclusion or generalization – Devise and implement a strategy in order to interpret, evaluate and validate a mathematical solution to a contextualized problem
Using symbolic, formal and technical language and operations:	– Use appropriate variables, symbols, diagrams, and standard models in order to represent a real-world problem using symbolic/formal language – Understand and utilize formal constructs based on definitions, rules, and formal systems as well as employing algorithms – Understand the relationship between the context of the problem and representation of the mathematical solution. – Use this understanding to help interpret the solution in context and gauge the feasibility and possible limitations of the solution
Using mathematical tools	– Use mathematical tools in order to recognize mathematical structures or to portray mathematical relationships. – Know about and be able to make appropriate use of various tools that may assist in implementing processes and procedures for determining mathematical solutions – Use mathematical tools to ascertain the reasonableness of a mathematical solution and any limits and constraints on that solution, given the context of the problem.

Source: OECD, 2019.

The capabilities enlisted in Table 1, for instance, reasoning and argument, formulating situations mathematically can be measured by arranging the questions' contents, and how the students approach the questions, which problem-solving strategies they use, or which misconceptions they have on the items can also be measured by involving open-ended questions (Lie, Taylor, & Harmon, 1996). In other words, arranging the question contents or diversifying the testing procedure (e.g., involving both multiple-choice and open-ended questions together) render measuring cognitive as well as the metacognitive performance of the students possible.

Of identifying metacognitive abilities, for instance, it has been shown that students' reading abilities are highly linked to their metacognitive abilities (Myers & Paris, 1978; see also, White & Frederiksen, 2005). Additionally, Ardelt, Shiefele, and Schnieder (2001) showed that metacognitive knowledge on understanding the material that is read is also closely related to the students' reading ability (i.e., test score) at PISA 2000 test. The same relation has also been observed in the proceeding administrations of PISA tests, such as at PISA 2009 (e.g., Ardelt & Schneider, 2015). As an alternative assessment method, a scenario-based metacognitive knowledge test has also been developed to be used for measuring the students' higher-order cognitive abilities such as learning and problem-solving strategies (Handel, Ardelt, & Weinert, 2013).

Despite the existence of the above-mentioned assessment methods, the investigations of cross-culturally administered PISA tests assess the students' metacognitive abilities in separate questionnaires while grading their cognitive performance right in the test (see also, Maag Merki, Ramseier, & Karlen, 2013; Wirth & Leutner, 2008). Therefore, there seems to be a lack of research that measures the students' metacognitive abilities without a need of administering any related scales along with the test. However, the study of Higham (2007) that investigated metacognitive abilities at Scholastic Aptitude Test (SAT)

emerges as an exception to this scarcity. In this study, Higham compared the specific SDT methods with the calculation methods suggested by Koriat and Goldsmith (1996).His findings confirmed that "regulation accuracy" referring to the withholding tendency of the respondents to answer to yield a higher number of reported correct answers seems to be measured more accurately by the SDT's methods compared to the alternative offered as a threshold model could not.

In this current study, it was aimed at using the calculation method offered by Higham (2007) specifically in the assessment of the PISA mathematics test for the first time. Hence, along with scoring the abilities at involved in PISA mathematics, it was targeted to detect metacognitive monitoring performance of the students, referring to one's ability to differentiate their correct responses from among the incorrect ones with the calculation methods of SDT detailed in the following section, and to reveal the relationship between this performance and the abilities (i.e., latent classes) that are to-be-obtained by the CDM's method.

There are numerous studies of research on "source monitoring" that reveals whether knowledge is produced in memory (false memory) or it is retrieved from a memory storage containing the true knowledge of an actual event (true memory) and on how aging affects the cognitive abilities (e.g., Baltes, Staundinger, & Lindenberger, 1999). However, these studies do not seem to directly investigate or study metacognitive monitoring performance as a variable. In other words, it not clear why some people are good at this ability yet others are not. In this regard, as was suggested by Dunlosky and Tauber (2001), metacognitive monitoring performance declines with age while aging is also observed with some declines in, for instance, short term memory and episodic memory performance, and in problem-solving abilities. However, it does not seem clear as to whether cognitive abilities decline due to gradually poorer metacognitive abilities observed by aging since declines in the cognitive and metacognitive abilities are measured simultaneously. The current study, however, differs from the existing research in the way that it investigates which sub-abilities that the monitoring performance might be closely linked and that it aims to classify the students' cognitive abilities with CDM and compares the

The present study, therefore, differs from the existing research that has been run particularly in the education field in the way that it aimed at revealing which (sub-)capabilities are linked to the metacognitive monitoring and to study how the students' metacognitive abilities differ with the hypothesis testing while classifying their cognitive capabilities by the CDM methods both in a single testing procedure and in the same test.

The following sections lay out the signal detection theory and cognitive diagnostic model as used in the current study.

Signal Detection Theory and Metacognitive Monitoring

Metacognition, defined as "the knowledge about knowledge" by Flavell (1979), has taken the interest of many researchers. The researchers can measure metacognitive judgments by various standard methods such as ease of learning (EOL), judgment of learning (JOL), feeling of knowing (FOK), etc. The common denominator of these studies and their results' main output is that they allow measuring how well individuals are aware of their knowledge objectively. In this vein, it seems critical for the studies that investigate cognitive abilities along with metacognitive performance to detect what level of awareness individuals have on their responses' correctness beyond studying how well respondents are good at knowing the correct answer per se (e.g., quantifying the number of correct answers given at a test). In other words, measuring the awareness of the respondents on their answers' correctness can provide the researchers with further parameters beyond counting the number of correct responses reported.

Though the measurement of metacognition can be applied by various methods (e.g., the methods to quantify the judgments given for the ease of learning, feeling or knowing, etc.), a particular method named Type-2 SDT, which is fundamentally based on the Green and Swets' (1966) Signal Detection Theory, provides a standard alternative for this measurement. Based on the basic assumptions

of the SDT, participants' metacognitive monitoring performance can be measured by calculating their "hit" and "false alarm" rates as well as their "response criteria", where the latest refers to how strictly or leniently one behaves when responding. Signal Detection Theory that is based on "psychophysics", referring to a field that focuses on the relation between physical stimulus and its sensational and perceptual reflections (Luce & Krumhansl, 1988), can be categorized into two as Type-1 and Type-2 SDT (e.g., Higham, 2002; Higham & Tam, 2005). In standard Type-1 SDT procedure, the respondent is asked to decide whether they detect the signal buried in, for instance, a white noise ("Yes"/"No"). The signal may exist as buried in the noise in some trials, yet it does in fact not exist in some other trials. According to the theory, it is assumed that the trials containing the signal and those having no signal at all construct two normal distributions and the participants tend to say "Yes" (i.e., "Yet! It exists") if the trial is above the response criterion set and says "No" (i.e., "No! It does not exist") if the trial is below this criterion. The probabilities of the participants' decisions given as "yes"/"no" can be calculated concerning the conditions where the signal is present or absent; see Table 2. According to this contingency, the *hit rate* refers to the number of "yes" decisions given out of the total number of the trials where the signal was present. The *false alarm rate*, however, is the number of "yes" decisions given by the participant out of the total number of trials where the signal was absent. The participant can also give a "no" decision correctly in the trials where the signal is absent. When the rate measured in such trials is called the *correct rejection rate*, the participant may also make a "no" decision in the trials where the signal is present. The latest one is called *miss rate* (Abdi, 2007). In short, this observation called Type-1 focuses on whether the participant who plays a machine-like role passively detects the existence of the signal correctly and this signal detection method is defined as a stimulus-contingent one (Higham & Tam, 2005).

Table 2. Four possible types of responses that can be obtained according to the Type-1 Signal Detection Theory

	Decision (participant's response)	
Reality	"Yes"	"No"
Signal present	Hit (a)	Miss (b)
Signal absent	False Alarm (c)	Correct rejection (d)

Note. hit rate = a/(a+b); false alarm rate = c/(c+d); miss rate = b/(a+b); correct rejection rate = d/(c+d).
Source: Abdi, 2007.

Despite that the rates' names are the same, their features in the Type-2 signal detection are different. In this type, for instance, the respondent is assumed to generate a list of candidate answers in their minds containing correct and incorrect answers. In other words, it is assumed that a generation of candidates executed is executed before reporting the answer instead of the knowledge (i.e., the response) is not directly retrieved from memory like a vacuum (e.g., Bahrick, 1970; Kintsch, 1970; see also, Watkins & Gardiner, 1979). Assuming that correct and incorrect candidates normally distribute in terms of their memory strengths which thereby vary their confidence levels concomitantly, Type-2 SDT aims at quantifying the respondent's ability to discriminate their correct answers from among incorrect ones by measuring the divergence between the mean values of these distributions (e.g., d'). The report criterion in Type-2 SDT, however, is defined as the criterion above which the participant reports the generated answers and below which they withhold the generated candidates. In other words, generated answers are tended to be reported if they are above the report criterion and withheld if they are below this criterion set. As inferred from this assumption, participants are asked to decide whether "they wish to answer" or "do not wish to answer the question" (e.g., "report" or "pass"). Participants are still asked to make their best guesses even if they prefer not to answer (i.e., "pass"). In these cases, for instance, the participant might have inserted a high report criterion so that might have chosen the "pass" option for a potentially correct answer. That is, it is expected for this participant that they tend to report only those answers on which they are highly confident on the responses' correctness. Lastly, participants are asked to rate their confidence on the correctness of their responses on a Likert-type scale. A contingency table for the type-2 SDT, in short, is as follows: 2 (response: reported vs. withheld) x 2 (candidate answer:

correct – incorrect) (see also, Higham & Tam, 2005, p. 599). As the distribution of the incorrect answers generated (Figure 1(b), the distribution on the left) starts with the lowest confidence level, the distribution of the correct answers generated (Figure 1(b), the distribution on the right) disperses with higher confidence levels. To summarize, one's ability to discriminate correct and incorrect responses is a function of how much further apart the distributions of correct and incorrect responses generated as well as at what level their report criteria are set (e.g., stringent or lenient). Therefore, Type-1 SDT is defined as a stimulus-contingent SDT whereas Type-2 SDT is a response-contingent one (Higham, 2002, 2011).

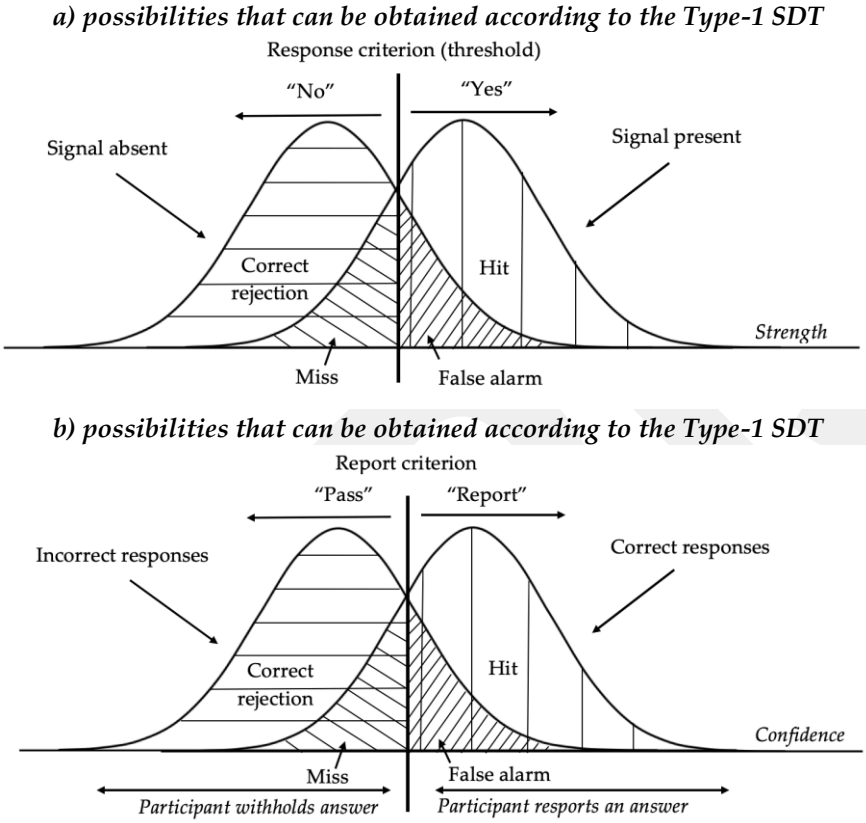

Figure 1. Probabilities that can be obtained according to Type-1 and Type-2 SDT (Source: Higham & Arnold, 2007)

It is possible to quantify the level of metacognitive monitoring ability after calculating the hit and false alarm rates emerge in the two distributions of the incorrect and correct responses generated. Researchers can calculate the distance between the means of these two distributions with, for instance, *d'* (*d prime*= Z[hit rate] - Z[false alarm rate]). This value can also be displayed on a graphic with the ROC (receiver operating characteristics) method. For this purpose, participant's hit and false alarm rates ranging between "0.00" and "1.00" at each cumulative confidence level are calculated and the intersections of these rates at the same cumulative confidence level are marked on a scatter plot; see Figure 2. The area emerging between the diagonal line that indicates a pure guess (i.e., indicates a reporting that can already be correct with a %50 chance) and the line connecting each intersection points can be calculated. The larger is the area as well as has a positive value, the higher is the area under the curve value of this participant (so that the *d'* value), which indicates one's ability to concordantly detect correct responses as correct and incorrect responses as incorrect. The values having a negative sign, however, indicates that the participant decides their correct responses as incorrect yet considers their incorrect responses as correct. Figure 2 displays a ROC of a hypothetical case that can be used to calcite their area under the curve value.

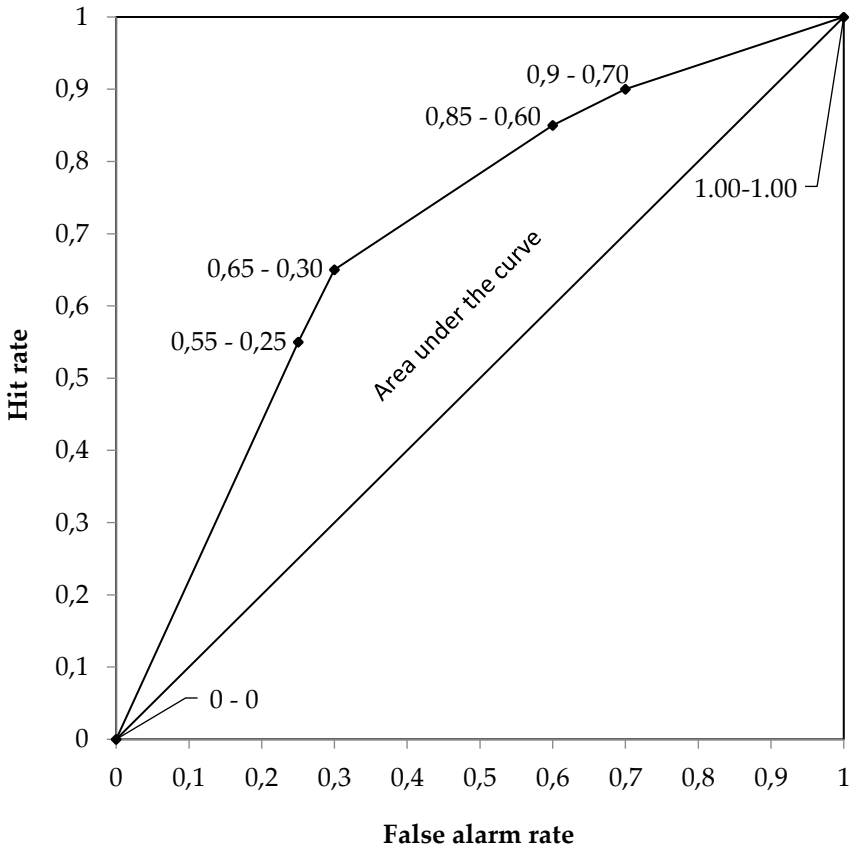

Figure 2. Receiver Operating Characteristics (ROC) that is drawn to calculate the area under the curve for a hypothetical case whose confidence ratings varied between "1" (not at all confident correct) and "5" (completely confident correct).

As displayed in Figure 2, the area under the curve is the area that appears between the broken line and the diagonal line on the graph. The highest intersection point on the broken line (1.00-1.00) is the intersection point of the hit and false rates calculated for the cumulative confidence level "1+" (1 and above) and the point counted as the 5th one from the top in the graph is the intersection point where the hit and false alarm rates calculated for the confidence level "5 only" (e.g., 0.55 hit & 0.25 false alarm rates). As can be inferred from the diagram, the broken line will be gradually smoother as the Likert-type rating has a higher number of anchors so that the monitoring performance value will be measured more precisely.

Cognitive Diagnostic Model

The CDM used in the study aimed at classifying the students accurately whether they have the measured ability or do not have this ability by calculating the item parameters and the a priori values of the features (i.e., abilities) that are measured in the test. The level of how well the features measured in the test are represented by the test items is defined when the test is being developed via using the cognitive diagnostic model's calculations, which are based on both the latent class analyses and the item-response theory (IRT) approach. The results of the analyses classify the students in terms of their abilities. In short, the main objective in these analyses is to accurately determine which latent classes the students are in (Leighton & Gierl, 2007; von Davier, 2014).

Deterministic Input Noisy and Gate (DINA) Model developed by Haertel (1989) is a latent class analysis just like many other Cognitive Diagnostic Models (e.g., Junker, 1999; Junker & Sijtsma, 2001; MacReady & Dayton, 1977). The model is based on the relation between the items and the features and the model must work well to accurately determine the features that need each item in the test is responded correctly (de la Torre & Chiu, 2015; de la Torre & Lee, 2010). Shortly, two latent classes for

a to-be-developed or already administered test can be determined by, for instance, "k" number of features. For instance, the respondents can be classified into eight latent classes in a test where only three features are measured.

The possible classes for these three classes can be ordered as "000", "100", "010", "001", "110", "011", "101", and "111". Whereas those who show no features at all are in the first-order class ("000"), those who show the first and the third features are in the seventh-order class ("101). The critical point in this analysis is that the decision whether a respondent has or does not have the designated feature is completely probabilistic. The decision whether a respondent (e.g., a student or learner) is in the class of "0" or "1" (i.e., having this defined feature or not, respectively) is a probability value and this value has conventionally ".50" threshold although it can be defined by the researcher. In other words, if the value of one's probability to have the given feature is below ".50", they are classified under class "0" (i.e., the student does not have this feature), and if this value is equal to or above the threshold, then the student is classified under the class "1" (i.e., the student has the defined feature) (see also, Başokçu, 2012, 2014; de la Torre & Douglas, 2004; de la Torre, 2008).

The mathematical abilities, being one of the independent variables in the present study, are taken as categorical rather than a continuous variable and the decisions whether the students had or did not have these features (i.e., abilities) are given by the DINA Model, which is one of the available methods of the CDM methods.

Method

Participants

The sample is composed of 6th-grade students who are enrolled in a public school in Bornova, Izmir province. Two-hundred-and-thirty students (110 male, 120 female) who constituted a totally seven same-grade sections in the school volunteered to participate in the study. The analyses on the test and the item parameters as well as the latent class analyses measured by the DINA model are analyzed with the data collected from these 230 students. The analyses on the monitoring performance, however, were analyzed among 130 students (58 male, 72 female) who were randomly selected from the total sample and who were administered monitoring procedure (e.g., they solved the test with "report/pass" option and then giving confidence ratings for their responses). Since 10 students handed in the test with a bulk of missing values, they were excluded from the AUC analyses. Therefore, AUC analyses were run on the data gathered from the remaining students (N=120). The number of students who were randomly selected and were administered with the monitoring procedures is homogeneous among the sections. The test sample and the experimental sample did not differ in terms of their test scores (M=2.41; SD=2.28; M=2.44; SD=1.83, respectively).

Procedure

The study was conducted on a pilot sample group of the general sample of the TUBITAK Project No. 115K531. The whole procedure was run in coordination with the Directorate of National Education of Izmir Province. The students were provided with the parent consent forms by the school administration and only those students whose parents gave their consents were taken to the study. A limited number of students whose consents were not taken were still in the classroom while the test was being administered and they were asked to do a reading activity meanwhile. The 12-item PISA-equivalent mathematical test, which is detailed in the "Materials" section, was simultaneously administered to totally 230 6th-grade students in their classrooms. The officers handed in the instructions and the test booklets to the project assistants, test administrators, and the invigilators. The data collected were posted to the project team by the officers after they inserted each filled-in test booklet in sealed envelopes.

Materials

The study utilized a 6th-grade level mathematics test that was composed of 12 items and aimed at measuring the mathematical abilities at the PISA test. The test development and the detection of measured abilities processes are implemented with the field experts in mathematics education.

Considering the item types, the subjects, and the grade level, the project researchers, advisors, and the teachers who had expertise in the fields together subsumed the PISA capabilities adapted for the study under four categories. The categories entitled "communication and association", "mathematization", "reasoning and developing strategy", and "using symbolic and technical language" are defined as follows (Başokçu, 2019):

Communication and association: The ability that covers associating the mathematical language with daily language and symbols and interpreting the accuracy and meaning of the mathematical ideas. It also covers associating mathematical concepts with each other, other disciplines, and with the real world.

Mathematization: It refers to the activities such as modeling, structurally presenting, identifying with assumptions, formulating a problem in mathematical form, gaining and interpreting mathematical outputs of an established construct or a model.

Reasoning and developing strategy: It is the process of gathering new information by using the specific mathematical tools (e.g., symbols, definitions, relations, etc.) and thinking methods (e.g., induction, deduction, comparison, generalization, etc.) with the information in hand. Developing strategy refers to selecting and designing a strategy to use mathematical knowledge and abilities in problem solving.

Using symbolic and technical language: In terms of the mathematics literacy, using the ability of symbolic and technical language covers the behaviors of understanding and interpreting the symbolic displays of the mathematical contents that are defined with the mathematical rules.

The item parameters, the above-mentioned mathematical abilities that were measured by the test items, and the DINA model item parameters are displayed in Table 3 (Table 3(a), 3(b), 3(c), respectively).

Table 3. Item parameters of the test material (a), mathematical abilities measured by the test items (b) (0=the item does not measure the defined ability; 1= the item measures the defined ability), and DINA model item parameters (guess & slip) and the error scores of these parameters (c)

a) Item parameters of the test material												
	Test items											
	1	2	3	4	5	6	7	8	9	10	11	12
Difficulty	0,28	0,12	0,02	0,37	0,33	0,11	0,24	0,07	0,35	0,28	0,05	0,18
Discrimination	0,46	0,26	0,06	0,58	0,47	0,27	0,37	0,16	0,58	0,42	0,12	0,35
Item-test correlation	0,47	0,51	0,51	0,53	0,45	0,68	0,48	0,55	0,55	0,48	0,65	0,5
b) Mathematical abilities measured by the test items												
Communication and association	1	1	0	1	1	0	0	0	0	1	1	1
Mathematization	0	0	1	0	0	1	1	1	1	0	0	0
Reasoning and developing strategy	1	1	1	0	0	0	1	1	1	0	1	0
Using symbolic and technical language	0	0	1	1	1	1	0	0	1	0	1	0
c) DINA Model item parameters and standard errors of the measurement instrument												
Guess (G-par)	0,21	0,04	0,02	0,16	0,33	0,07	0,12	0,04	0,31	0,10	0,03	0,17
Guess standard error	0,04	0,03	0,01	0,05	0,04	0,03	0,05	0,02	0,04	0,06	0,01	0,05
Slip (S-par)	0,21	0,25	0,54	0,06	0,26	0,46	0,41	0,45	0,38	0,24	0,49	0,32
Slip standard error	0,10	0,10	0,05	0,08	0,07	0,11	0,14	0,06	0,13	0,07	0,07	0,05

The item difficulty mean was 0.20, the discriminability mean was 0.34, and the mean of item-test correlations was 0.53; see Table 3(a). The test's KR-20 reliability coefficient was 0.73. This value indicates high reliability for a 12-item test. Test's validity and reliability were measured by the Item Response Theory (IRT) analyses. The analyses were run by 2-Parameter Logistic Model. The test information function and the item characteristics curves that displayed ogive functions were examined, and they indicated that the developed test measured the variable that the test intended to measure with high discriminability and validity. Additionally, once the multi-dimensional feature of the DINA model is considered, it was shown that the Classical Test Theory (CTT) and IRT evidence given for the test discriminability and validity should be used as a comparison criterion.

The process of matching (i.e., relating) items with the measured abilities are determined by considering expert opinions, and the related Q matrix was constructed as shown in Table 3(b). This process resembles the theoretical approach that matches the items with the dimensions in the test development process. However, the difference of CDM is that any single item can be matched with more than one feature (i.e., dimension) in the Q matrix. This matrix displays the fundamental a priori and theoretical relations that are used for the CDM analyses. DINA model item parameters and the latent classes of the respondents who took the test can be identified by using this matrix. As displayed in Table 3(b), seven items measured the "communication and association", five items measured the "mathematization", seven items measured the "reasoning and developing strategy", and six items measured "using the symbolic and technical language" abilities. The fact that one item can be matched with more a single feature, which is the distinctive feature of CDM, is shown in Table 3(b). In other words, whereas only two items measured "one" feature, seven items measured "two" features, and three items measured "three" features, which is the depiction that shows the model's multi-dimensional nature. As seen in Table 3(c), "guess" parameters (G-par) are low. "Slip" parameters (S-par), however, increase particularly with the items that have higher difficulty values. While this result is something expected for CBM, it seems the model fits well once the general scores and their means are considered. Also, the relative indexes for the model's fitness were calculated for the Akaike and Bayesian Information Criteria (AIC & BIC) as 2535.38 and 2669.46, respectively.

The participants' metacognitive monitoring performance was scored individually by using the area under the curve (AUC) method, shown in Figure 2. The participants were asked to read the questions first and then choose one of the following options for each item: "I believe I can solve the question correctly" vs. "I believe I can't solve the question correctly". After solving the questions (and after making their best guesses even if they passed the questions), the participants also rated their confidence levels on the correctness of their answers on a 5-anchor Likert-type scale (1= "not at all confident correct"; 5= "completely confident correct"). All hit and false alarm rates at each cumulative confidence levels (i.e., "1 & above", "2 & above", "3 & above", "4 & above", and "5 only" confidence levels) were calculated for each of the participants individually and then their ROC curves were drawn, by which the area under the curve appearing under this ROC curve and the diagonal line shown in Figure 2 could be measured. For instance, the hit rate at "1 and above" confidence level is the rate of the number of correct responses reported by the participant out of the total number of correct responses that were reported and passed together no matter which confidence level they were rated with. However, when calculating the hit rate at "2 and above" confidence level, the correct responses which were rated with a confidence level of "2" or "any level above this" were taken into the calculation. On the other hand, the false alarm rate at "1 and above" confidence level is the rate of the number of incorrect responses reported out of the total number of incorrect responses that were reported or passed no matter which confidence level they were rated with. Likewise, when calculating this rate at "2 and above" confidence level, only those incorrect responses which were rated with a confidence level of 2 or any other above this level were considered (see also, Higham & Tam, 2005). As a result, each hit and

false alarm rates at each cumulative confidence were calculated for each of the participants separately via the above-mentioned formulae, and the sizes of the AUCs calculated were defined as the participants' metacognitive monitoring performance (i.e., the ability to discriminate correct and incorrect responses).

Results

The data gathered in this study that investigated the mathematical abilities of the 6th-graders at a PISA mathematics test were analyzed by the latent class analyses based on the DINA model and have not been used in evaluating this test so far. The CDM packages developed for the Ox Edit and R programs were used in the analyses (George, Robitzsch, Kiefer, Groß, & Ünlü, 2016). The metacognitive monitoring scores, however, were obtained by the Type-2 signal detection theory's calculation methods that have been used only in the SAT so far (Higham, 2007) and the area under the curve calculations (see also, "Signal detection theory and metacognitive monitoring" section and Figure 1). The posterior probabilities of the participants' latent classes, the observed probabilities of the classes, the mean correct answers in the latent classes, and their metacognitive monitoring scores are shown in Table 4.

Table 4. The posterior probabilities belonging to the latent classes classified with DINA, the means of correct answers in the latent classes, and the metacognitive monitoring scores of these classes measured with area under the curve (AUC) calculations

Latent class*	Posterior probability	Observed probability	Correct (mean)	Metacognitive monitoring ability (AUC)
"0000"	0,0658	0,187	1,12	0,16
"1000"	0,0631	0,106	2,50	0,14
"0100"	0,0658			
"0010"	0,0658			
"0001"	0,0658	0,132		
"1100"	0,0631	0,081	1,75	0,13
"1010"	0,0538	0,074	2,50	0,18
"1001"	0,0644	0,076	2,21	0,13
"0110"	0,0572	0,052	2,21	0,22
"0101"	0,0623			
"0011"	0,0658			
"1110"	0,0476	0,084	4,67	0,34
"1101"	0,0527	0,087	3,89	0,34
"1011"	0,065			
"0111"	0,0482			
"1111"	0,0936	0,121	6,36	0,41

* Latent classes ("XXXX") are displayed with four digits and these digits refer to "communication and association", "mathematization", "reasoning and devising strategies", and "using symbolic and technical language", respectively. The digits "1" and "0" indicate whether the designated capability exists or does not exist, respectively. For instance, the "1010" code indicates that this latent class coded as "1010" has the abilities of "communication and association" and "reasoning and devising strategies" both yet it does not show the abilities of "mathematization" and "using symbolic and technical language".

The posterior probability values displayed in Table 4 are the values that belong to the values of the whole latent classes that are calculated in terms of the model's parameters. However, not all of the latent classes can be observed in every sample. As seen in Table 4, only nine latent classes could be observed in the sample out of totally 16 possible latent classes. The posterior probabilities of the latent classes are close to each other as again seen in Table 4. This is accepted as an indicator that shows the defined features (i.e., abilities) are independent of each other (Chen, de la Torre, & Zhang, 2013; Huo & de la Torre, 2014). As seen in Table 4 again, the higher is the number of features observed in the latent

classes (in other words, the total number of "1" showing the defined feature is involved in the class increase), the higher are the test means as well as the monitoring scores of the latent classes. For instance, the latent class coded "0000" which showed no ability in the test had 1.12 correct scores on average whereas the latent class coded "1111" which showed all of the abilities in the test obtained a mean of 6.36 correct responses from the same test. The monitoring performance of the group (i.e., the latent class) who showed no abilities in the test was significantly (.16) lower than the group who showed all of the abilities (.41); $t(50)=3,59$, $p<0,001$. A series of independent t-test analyses were run between the conditions of having the ability or not having this given ability to detect which features differ in terms of metacognitive monitoring performance; see Table 5.

Table 5. Independent-samples t-test results when the groups that either have the designated ability (1) or not (0) are compared in terms of their metacognitive monitoring performance (AUC scores)

	Whether having the ability or not	*n*	*M*	*s*	*t*	*df*	*p*
Communication and association	1	60	,23	,29	,953	118	,343
	0	60	,18	,28			
Mathematization	1	40	,28	,27	1,823	118	,071
	0	80	,17	,29			
Reasoning and developing strategies	1	42	,30	,30	2,914	118	,004
	0	78	,15	,26			
Using symbolic and technical language	1	44	,23	,31	,945	118	,347
	0	76	,18	,27			

Note. *n*=number of the participants; *M*=means of the monitoring scores; *s*=standard deviation, *df*= degrees of freedom; *p*=alpha value

According to the t-test results displayed in Table 5, having the ability of "reasoning and devising strategies" yielded higher monitoring performance than those who did not show this ability. In other words, the mean of the monitoring scores of those students showing reasoning and devising strategies ability is significantly higher (.30) than those students who did not show this ability (.15) with a medium-size effect size (Cohen's *d*=.55). However, the students did not differ in terms of their monitoring scores concerning whether they showed the abilities of communication and association, mathematization, and using symbolic and technical language.

Discussion, Conclusion and Suggestions

CDM analyses allocate students to the latent classes in terms of relating the test items to the features measured by the test by considering the responses given (Henson & Douglas, 2005). Structuring the test and its features is also conducive to lay out a measurement model for the designated latent classes (DiBello, Roussos, & Stout, 2006). Therefore, CDM parameters inform the researchers regarding the model's fitness (Hu, Miller, Huggins-Manley, & Chen, 2016). Based on the findings of the current research, the CDM parameters and the posterior probabilities of the latent classes indicate that the fitness level of the measurement model constructed by considering mathematical abilities is valid. This finding exhibits the validity of test items (shown in Table 3(a)) and the Q matrix (shown in Table 3(b)). The psychometric properties of the administered test were studied with the classical test theory as well as the item response theory and the analyses run by both methods provided evidence regarding the test's reliability and the items' validity. Once these findings are considered, it seems that the findings on the classification determined by the DINA model were highly valid measurements.

The second dimension of the study investigated the relationship between metacognitive monitoring, which refers to one's awareness of the correctness of their responses, and mathematical abilities. In many studies of research, metacognition is contained under the topic of higher-order thinking abilities (e.g., Brookhart, 2010; Conklin, 2012; Schraw & Robinson, 2011; Williams, 2003). The

monitoring performance of those who showed the "reasoning and developing strategy" ability in particular among the four abilities determined in the study and those who did not show this specific ability differed significantly from each other. Similar results were also obtained in previous research (see also, Kramarski & Mevarech, 2003; Schneider & Ardelt, 2010). For instance, one study of Kramarski and Maverech (2003) who studied 8th-graders showed that the students who were thought with cooperative learning together with the metacognitive training and the students who were thought with individual learning along with the metacognitive training were significantly more successful at graphic construction as well as at metacognitive knowledge performance than those who studied the same material with the individual learning.

The current finding that the ability of reasoning and developing strategy is highly related to the metacognitive monitoring performance seems in line with, for instance, the model of Koriat and Goldsmith (1996) which has been utilized by gradually more researchers in the literature. According to this framework model, entitled "strategical regulation of memory accuracy", Koriat and Goldsmith suggest the following. Participants may develop a metacognitive strategy to render their responses be composed of correct ones only when a free-report method is used to answer the questions at a given test (that is, the participants are asked to report all of the correct answers only that they know). For this strategy, let us consider a participant who is taken to a free-report test after studying a 20-item word list and, say, this participant, tagged as participant A, reports 12 words only. Say, again, eight words among these 12 words reported are remembered and reported correctly. Another participant, now tagged as participant B, reports eight words for the same studied list; however, say, six of the words out of these eight words reported are correct. According to Koriat and Goldsmith, despite that the participant B might have performed a lower "quantity performance" than the participant A (in other words, reported a lower number of correct responses [6] out of the total number of items studied [20]), the participant B could strategically increase the level of their "accuracy performance". That is, the participant B could increase their accuracy performance by reporting less. The "quantity performance" that can be calculated easily with the model is the rate of the number of correct responses out of the total number of all possible correct responses (for this example, it is the total number of the words in the list). Therefore, the quantity performance of participant A is .40 (8/20) and of participant B is .30 (6/20). The accuracy performance of participant A, which is calculated as the rate of the total number of correct responses out the total number of responses reported, is .66 (8/12) and of the participant B is .75 (6/8). In short, one of the participants (i.e., the participant B) develops a metacognitive strategy by reporting fewer responses in the free-report test and implementing this strategy via inserting a higher level of response criterion (i.e., not reporting any word that comes to their minds regardless of whether it is a correct or an incorrect one) and/or not recognizing the correctness of the remembered words well enough as the other participant. Therefore, despite having a lower number of words that were reported, this participant can regulate their memory accuracy "by reporting a higher number of correct answers among all of the answers reported". It seems herein so critical that the present study showed the ability of "reasoning and developing strategy" is significantly related to the metacognitive monitoring performance, which is also parallel to the Koriat and Goldsmith's model in the way that this very ability is a regulation activity implemented strategically and is completely a high-order memory performance. The findings of the study showed that the ability of reasoning and developing strategy rather than the abilities of communication and association, mathematization, and using symbolic and technical language are directly and significantly linked with one's ability to discriminate correct and incorrect responses, seeming parallel to the Koriat and Goldsmith's arguments. In this vein, the findings on which ability or abilities that were detected by CDM are particularly related to the monitoring performance that was calculated with the SDT in the current study seems in line with the model of Koriat and Goldsmith on the strategical regulation of memory accuracy.

Although there may exist a hot debate between Higham (2002) and Koriat and Goldsmit (1996) regarding which calculations are better at measuring the metacognitive monitoring performance (see also, Higham, 2011), the findings of the current study imply that the metacognitive monitoring and strategical regulation are highly interlinked abilities. Despite the arguments on which method, type-2 SDT suggested by Higham or the method proposed by Koriat and Goldsmith, would be better at measuring the "memory accuracy", both of the methods aim at measuring basically the same ability. The uniqueness of the present study, however, appears in the way that it reveals the latent abilities that can be closely related to the monitoring ability, calculated by the type-2 SDT method of Higham (2002), via detecting these possible abilities measured by the questions' contents (e.g., "reasoning and developing strategy") and with the help of latent class analysis (e.g., CDM). The question, however, asking why showing a better ability of reasoning and developing strategy is not resulted by having a high monitoring ability may be raised. As a response to this possible question and based on the findings and the implications we gathered in this study, we believe that the metacognitive monitoring ability is an ability that is composed of various sub-abilities. Therefore, it is highly likely that the ability to discriminate correct responses from among the incorrect ones is possibly acquired only after one has the abilities of reasoning and/or developing a memory strategy as sub-abilities. The first reason for this proposal is that metacognitive monitoring seems to be considered as a holistic (i.e., a general) ability in the related literature. Additionally, it is not clear as to why some people are good at this ability yet some others are not. The research that investigates the cognitive changes with aging, in particular, appear as mainly the correlational studies, unlike the current study that used hypothesis testing. To better answer the above-mentioned question with different experimental designs, the future research may, for instance, consider comparing the monitoring performance of the students who are trained to gain the ability to develop strategy with a control group of students which thereby could provide evidence on the direct effect of developing strategy ability on monitoring. It also seems critical for the prospective research to reveal various other sub-abilities that might be related to the monitoring performance and investigate and to investigate monitoring ability at other PISA tests, such as science and language, and among other age groups.

Besides the above-mentioned suggestions, the current study has some limitations. First, the study was planned to be run with a test that was developed to measure higher-order thinking abilities of the 6th-graders at mathematics. Therefore, conducting the same study with a similar test that should now be developed for more fundamental features is important. The comparison of different approaches existing for the CDM models is also critical to expand the related literature. It can also be stated that that increasing the sample size, rearranging the test duration, varying the subjects and the grade levels in such a study that simultaneously used two simultaneous, which were relatively recent for cognitive psychology and psychometry, may contribute to the literature even further.

To summarize, the findings of the current study provide a piece of cross-validation evidence on the validity of the classification model established by the CDM. It is not only the statistical evidence but also the rational evidence that is needed to be sought after to validate the models constructed particularly in the studies that use the latent class analyses. The current findings reveal rare empirical evidence on the rational validity of the CDM classifications. Therefore, it is believed that this study would potentially be a reference study for prospective research on this subject. Lastly, because CDM models are based on the fitness between the model and the data and the studies on the measurement validity are based on relative criteria, which thereby reveal some results that cannot be assessed with an absolute criterion, it is herein proposed that the models' utilization fields could be expanded once the psychological construct and the psychometrical findings support each other in the model-and-data fitness.

References

Abdi, H. (2007). Signal detection theory. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics.* New York: Elsevier.

Albacete, R. O., Cascón, J. A., Arnal, L. A., Pérez, J. D., Domínguez, I. E., & Sánchez, F. S. (2016). Psychometric properties of the reading comprehension test ECOMPLEC. Sec. *Psicothema, 28*(1), 89-95. doi:10.7334/psicothema2015.92

Ardelt, C., & Schneider, W. (2015). Cross-country generalizability of the role metacognitive knowledge for students' strategy use and reading competence. *Teachers College Record*, *117*(1), 1-32.

Ardelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Education, 16*, 363-383. doi:10.1007/BF03173188

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review, 77*, 215-222.

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology, 50*(1), 471-507.

Başokçu, T. O. (2012). DINA model parametreleri kullanılarak tahminlenen madde ayırıcılık indekslerinin incelenmesi. *Eğitim ve Bilim, 37*(163), 310-321.

Başokçu, T. O. (2014). Öğrenci yeteneklerinin kestirilmesinde bilişsel tanı modelleri ve uygulamaları. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 14*(1), 1-32. doi:10.17240/aibuefd.2014.14.1-5000091500

Başokçu, T. O. (2019). *A recommended model to increase success level of turkey in mathematics in international wide-scale exams. Effectiveness of the cognitive diagnosis based tracking model*. Izmir: TUBITAK 115K531.

Bean, J. C., & Peterson, D. (1998). Grading classroom participation. *New Directions for Teaching & Learning, 74*, 33-40. doi:10.1002/tl.7403

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Danver, MA: ASCD Publications. Retrieved from https://books.google.com.tr/books?id=AFIxeGsV6SMC

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140. doi:10.1111/j.1745-3984.2012.00185.x

Conklin, W. (2012). *Strategies for developing higher-order thinking skills*. CA: Shell Education.

de la Torre, J. (2008). An empirically based method of q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343-362. doi:10.1111/j.1745-3984.2008.00069.x

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical q-matrix validation. *Psychometrika*, 1-21. doi:10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*(1), 115-127. doi:10.1111/j.1745-3984.2009.00102.x

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). A review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979-1030). Hollanda: Elsevier. doi:10.1016/S0169-7161(06)26031-0

Dunlosky, J., & Tauber, S. (2001). *The Oxford handbook of metamemory*. NY: Oxford University Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist, 34*(10), 906-911. doi:10.1037/0003-066X.34.10.906

George, A., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1-24. doi:10.18637/jss.v074.i02

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*(4), 318-341. doi:10.1080/15305058.2010.509554

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Güzel, M. A., & Higham, P. A. (2013). Dissociating early- and late-selection processes in recall: The mixed blessing of categorized study lists. *Memory & Cognition, 41*, 683-697. doi:10.3758/s13421-012-0292-3

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301-321. doi:10.1111/j.1745-3984.1989.tb00336.x

Handel, M., Ardelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for Educational Research Online, 5*(2), 162-188. Retrieved from https://www.pedocs.de/frontdoor.php?source_opus=5481

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277. doi:10.1177/0146621604272623

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30*(1), 67-80. doi:10.3758/BF03195266

Higham, P. A. (2007). No special K! A Signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136*, 1-22. doi:10.1037/0096- 3445.136.1.1

Higham, P. A. (2011). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A. Higham & J. P. Leboe (Ed.), C*onstructions of remembering and metacognition. Essays in honour of Bruce Whittlesea* (pp. 109-127). Basingstoke: Palgrave Macmillan. doi:10.1057/9780230305281_9

Higham, P. A., & Arnold, M. M. (2007). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. (In special issue on Bridging cognitive science and education: learning, memory, and metacognition.). *European Journal of Cognitive Psychology*, *19*, 718-742. doi:10.1080/09541440701326121

Higham, P. A., & Gerrard, C. (2005). Not all errors are created equal: metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology, 59*, 28-34. doi:10.1037/h0087457

Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language*, *52*(4), 595-617. doi:10.1016/j.jml.2005.01.015

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*(2), 119-141. doi:10.1080/15305058.2015.1133627

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*(6), 464-485. doi:10.1177/0146621614533986

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Committee on the Foundations of Assessment*. Retrieved from http://www.stat.cmu.edu/~brian/nrc/cfa/documents/final.pdf

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 25*, 258-272. doi:10.1177/01466210122032064

Karakelle, S., & Saraç, S. (2010). Üst biliş hakkında bir gözden geçirme: Üstbiliş çalışmaları mı yoksa üst bilişsel yaklaşım mı?. *Türk Psikoloji Yazıları, 13*(26), 45-60.

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Normal (Ed.), *Models of human memory* (pp. 333-370). New York: Academic Press.

Koriat, A., & Goldsmith., M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490-517. doi:10.1037/0033-295X.103.3.49

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American Educational Research Journal*, *40*(1), 281-310. doi:10.3102/00028312040001281

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Lie, S., Taylor, A., & Harmon, M. (1996). Scoring techniques and criteria. In M. O. Martin & D. L. Kelly (Ed.), *Third International Mathematics and Science Study (TIMSS) technical report. Volume I: Design and development*. Chestnut Hill, MA: Boston College.

Lindblom-Ylanne, S., Pihlajamarki, H., & Kotkas, T. (2006). Self, peer, and teacher assessment of student essays. *Active Learning in Higher Education, 7*(1), 51-62. doi:10.1177/1469787406061148

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (Vol. 1, 2nd ed., pp. 3–74). New York: Wiley.

Maag Merki, K., Ramseier, E., & Karlen, Y. (2013). Reliability and validity analyses of a newly developed test to assess learning strategy knowledge. *Journal of Cognitive Education and Psychology, 12*(3), 391-408. doi:10.1891/1945-8959.12.3.391

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99-120. doi:10.2307/1164802

Myers, M., & Paris, S. G. (1978). Children's metacognitive knowledge about reading. *Journal of Educational Psychology, 70*(5), 680-690. doi:10.1037/0022-0663.70.5.680

OECD. (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem soloving knowledge and skills*. Paris: OECD. doi:10.1787/9789264101739-en

OECD. (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science* (Vol. 1). Paris: OECD. doi:10.1787/9789264091450-en

OECD. (2019). *PISA 2018 assessment and analytical framework*. Paris: OECD Publishing. doi:10.1787/b25efab8-en

Pieschel, S. (2009). Metacognitive calibration - an extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3-31. doi:10.1007/s11409-008-9030-4

Schneider, W., & Ardelt, C. (2010). Metacognition and mathematics education. *ZDM, 42*(2), 149-161. doi:10.1007/s11858-010-0240-2

Schraw, G., & Robinson, D. H. (2011). *Assessment of higher order thinking skills*. Kuzey Carolina: Information Age Pub. Retrieved from https://books.google.com.tr/books?id=6wAoDwAAQBAJ

Stacey, K., & Turner, R. (2014). *Assessing mathematical literacy: The PISA experience*. Cham: Springer International Publishing.

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 49-71. doi:10.1111/bmsp.12003

Watkins, M. J., & Gardiner, J. M. (1979). An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning & Verbal Behavior, 18*(6), 687-704. doi:10.1016/S0022-5371(79)90397-9

White, B., & Frederiksen, J. (2005). A theoretical framework and approach for fostering metacognitive development. *Educational Psychologist, 40*, 211-223. doi:10.1207/s15326985ep4004_3

Williams, R. B. (2003). *Higher order thinking skills: Challenging all students to achieve*. CA: SAGE Publications.

Wirth, J., & Leutner, L. (2008). Self regulated learning as a competence. Implication of theoretical models for assessment methods. *Journal of Psychology, 216*(2), 102-110. doi:10.1027/0044-3409.216.2.102

Wragg, E. C. (2001). *Questioning in the secondary school*. GBR London: Routledge.