Document generated on 04/25/2024 10:53 a.m.

Mesure et évaluation en éducation

Cognitive Diagnostic Analyses of the Progress in International Reading Literacy Study (PIRLS) 2011 Results

Dan Thanh Duong Thi and Nathalie Loye

Volume 42, Number spécial, 2019

Translation Issue

URI: https://id.erudit.org/iderudit/1084131ar
DOI: https://doi.org/10.7202/1084131ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Duong Thi, D. T. & Loye, N. (2019). Cognitive Diagnostic Analyses of the Progress in International Reading Literacy Study (PIRLS) 2011 Results. *Mesure et évaluation en éducation*, *42*(spécial), 127–165. https://doi.org/10.7202/1084131ar

Article abstract

Despite the grand demand to receive diagnostic information about students' difficulties in reading, there are very few tests specifically designed for diagnostic purposes. Therefore, many researches in cognitive diagnostic approach (CDA) use large-scale test results to provide fine and reliable diagnostic feedback on the strengths and weaknesses of students other than the total scores or percentiles ranks, which allow appropriate intervention. This study shows an example of the application of diagnostic modeling using data from 4,762 Canadian students who completed booklet 13 of the PIRLS test in 2011. The results highlight the potential for detailed diagnostic feedback of students' strengths and weaknesses on the underlying skills identified in the test.

Tous droits réservés © ADMEE-Canada - Université Laval, 2020

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

Cognitive Diagnostic Analyses of the Progress in International Reading Literacy Study (PIRLS) 2011 Results*

Dan Thanh Duong Thi
Université du Québec à Montréal

Nathalie Loye
Université de Montréal

Key words: cognitive diagnostic approach (CDA), reading, DINA, G-DINA, diagnostic classification models (DCM), large-scale tests

Despite the grand demand to receive diagnostic information about students' difficulties in reading, there are very few tests specifically designed for diagnostic purposes. Therefore, many researches in cognitive diagnostic approach (CDA) use large-scale test results to provide fine and reliable diagnostic feedback on the strengths and weaknesses of students other than the total scores or percentiles ranks, which allow appropriate intervention. This study shows an example of the application of diagnostic modeling using data from 4,762 Canadian students who completed booklet 13 of the PIRLS test in 2011. The results highlight the potential for detailed diagnostic feedback of students' strengths and weaknesses on the underlying skills identified in the test.

* French version : *Analyses diagnostiques cognitives des résultats du test du Programme international de recherche en lecture scolaire (PIRLS) 2011* – vol. 42, n°3, 29-70

Mots clés : approche diagnostique cognitive (ADC), lecture, DINA, G-DINA, modèles de classification diagnostique (MCD), épreuves à grande échelle

Malgré une importante demande de recevoir des informations diagnostiques sur les difficultés en lecture des élèves, il existe très peu d'outils d'évaluation conçus spécifiquement pour cet usage. Plusieurs recherches en approche diagnostique cognitive (ADC) utilisent donc les résultats d'épreuves à grande échelle pour fournir de la rétroaction diagnostique fine et fiable sur les forces et les faiblesses des élèves. Les modélisations de données permettent de s'éloigner des scores ou des rangs percentiles habituellement obtenus, et de fournir des pistes d'intervention appropriées. Cette étude vise à vérifier la faisabilité d'appliquer des modélisations à visée diagnostique aux résultats de 4762 élèves canadiens ayant fait le cahier 13 du test du PIRLS de 2011. Les résultats suggèrent un potentiel de recevoir de la rétroaction diagnostique détaillée de leurs forces et faiblesses sur les habiletés sous- jacentes du test.

Palavras-chave: abordagem diagnóstica cognitiva (ADC), leitura, DINA, G-DINA, modelos de classificação diagnóstica (MCD), testes em larga escala

Apesar da importante procura por informações diagnósticas sobre as dificuldades de leitura dos alunos, existem muito poucas ferramentas de avaliação concebidas especificamente para este uso. Diversas investigações em abordagem de diagnóstica cognitiva (ADC) utilizam, portanto, os resultados de testes em larga escala para fornecer feedback diagnóstico detalhado e fiável sobre os pontos fortes e fracos dos alunos. As modelizações de dados torna possível afastar-se das pontuações ou dos níveis percentuais normalmente obtidos e fornecer pistas de intervenção apropriadas. Este estudo tem como objetivo verificar a viabilidade da aplicação da modelização diagnóstica aos resultados de 4.762 alunos canadianos que realizaram o caderno 13 do teste do PIRLS de 2011. Os resultados realçam o potencial para um feedback diagnóstico detalhado dos pontos fortes e fracos dos alunos em relação às habilidades subjacentes ao teste.

Authors' note: Correspondence related to this article may be addressed to [duong_thi. dan_thanh@uqam.ca] and [nathalie.loye@umontreal.ca].

Introduction

There has been a significant increase in demand to diagnose student learning difficulties in recent years (de la Torre, 2009; Jang, 2009), but very few tests have been designed specifically for this purpose (Alderson, 2010; Jang, 2009; Lee and Sawaki, 2009; Leighton and Gierl, 2007). Several studies use the cognitive diagnostic approach (CDA) to analyse results from large-scale standardized tests such as the Test of English as a Foreign Language (TOEFL), the TOEFL Internet-based Test (iBT) or the Michigan English Language Assessment Battery (MELAB) to obtain detailed and reliable diagnostic feedback on students' strengths and weaknesses, which makes it possible to identify appropriate intervention (Gierl, Cui & Hunka, 2008; Hartz, 2002; Jang, 2005; Templin & Henson, 2006). Other studies (Dogan & Tatsuoka, 2008; Im & Park, 2010; Lee, Park & Taylan, 2011; Toker & Green, 2012; Lee, Park, Sachdeva, Zhang & Waldman, 2013; Arican a& Sen, 2015; Yamaguchi & Okada, 2018) are interested in analyzing data from large-scale international tests, such as the Trends in International Mathematics and Science Study (TIMSS), with diagnostic classification models (DCMs) based on the Q-matrix, that connects assessment items to knowledge or skills. These studies therefore provide detailed information on students' mastery of skills, on the link between teaching and student performance, and on countries' educational systems and their curriculum (Arican & Sen, 2015).

In the field of languages, research on this approach shows how reading competence can be decomposed into a set of knowledge and skills that can be diagnosed by psychometric modelling (Jang, 2005; Leighton & Gierl, 2007). However, so far, research in languages has only used tests administered to adults such as the TOEFL, TOEFL iBT or MELAB. Few studies have chosen a reading test developed for elementary school students, the age group most in need of intervention, given how this skill mastery influences their future academic success (Desrosiers & Tétreault, 2012; Pagani, Fitzpatrick, Belleau & Janosz, 2011).

Similar to TIMSS, the Progress in International Reading Literacy Study (PIRLS) develops a test to identify trends in reading performance among grade four students (Labrecque, Chuy, Brochu & Houme, 2012). Although the PIRLS test results provide information about student performance on two two purposes of readingand four reading processes, the Council of Ministers of Education, Canada, (CMEC) does not report individual student results (Labrecque et al., 2012). These results are used primarily for research purposes, which limits their exploitation and use to improve the teaching and learning of reading in elementary school.

The English test of PIRLS 2011 contains a total of 10 passages, including 5 passages of literary experience and 5 passages for reading to acquire and use information. Six passages are from previous cohorts' tests, while four passages were newly developed for the 2011 test (Labrecque et al., 2012). In total, there are 135 items, with 13 to 16 items for each passage that are almost equally divided between multiple-choice and short-answer questions (Labrecque et al., 2012). The passages and items were distributed into 10 blocks of 40 minutes each and then systemically organized into 13 booklets. The 13 booklets are therefore distinguished by the combination of text passages and associated items. Thus, this systematic combination ensures a balanced distribution of the type of text, the types of questions and the number of questions per objective and reading process. It therefore ensures equivalence of content among the different test booklets.

The content of only 10% of the passages and items was published for each cohort, allowing researchers to access these for study. Among the booklets of PIRLS 2011, booklet 13 had the most participants, 20.7% of students, while the other booklets each included only 6.6% of students. Due to the larger number of participants in booklet 13 and the possibility to access to item content, text, and student responses, we decided to analyze data from 4,762 Canadian students who completed this booklet.

The purpose of this article is therefore to verify the feasibility of applying diagnostic classification models to analyze the results of Canadian students who completed booklet 13 of PIRLS 2011. More specifically, we aim three objectives: 1) to evaluate the models fit to the data with the two Q-matrices, 2) to evaluate the diagnostic quality of the PIRLS test items, and 3) to examine the mastery or non-mastery of skills profiles of 4,762 Canadian students who completed booklet 13.

Literature Review

Link between the PIRLS test and theoretical reading models

PIRLS 2011 defines reading comprehension as the ability to understand and use those written language forms required by society and valued by the individual (Labrecque et al., 2012; Mullis, Martin, Kennedy, Trong & Sainsbury, 2009). This definition is based on theories that consider reading as a "constructive and interactive" process (Alexander & Jetton, 2000; Anderson & Pearson, 1984; Labrecque et al., 2012; Mullis et al., 2009). Students are "active constructors" of meaning because they engage effective cognitive and metacognitive strategies, as well as linguistic skills and background knowledge, to solve required tasks (Afflerbach & Cho, 2009; Anderson & Pearson, 1984; Carrell, 1983; Clay, 1991; Langer, 1990). In addition, reading is carried out in the interaction between text and reader in a particular context that promotes the reader's engagement so as to meet specific needs (Giasson, 1996; Grabe, 1991; Irwin, 1991; Mullis et al., 2009; Snow, 2002).

This interactive and constructive view of reading refers to interactive models (Rumelhart, 1978), the schema theory (Anderson & Pearson, 1984; Carrell, 1983), the construction-integration model (Kintsch, 1988) and the contemporary model in reading (Giasson, 1996; Irwin, 1991). Interactive models (Rumelhart, 1978) distinguish between two types of interactions: between reader and text and between skills and different types of knowledge (Carrell, 1983; Dubin, Eskey, Grabe & Savignon, 1986; Samuels & Kamil, 1984). In the first type of interaction, the reader's knowledge, cognitive and metacognitive strategies, physical and motivational characteristics, as well as the interaction of these characteristics influence the outcome of the reading process (Alderson, 2005). As for the textual elements, the content, the types of texts, the textual organization and the structure of the sentences, the typography, the relation between the verbal and non-verbal features as well as text presentation facilitate the comprehension of the text (Carell, Devine & Eskey, 1988; Grabe, 1991).

The interaction between the knowledge and skills used in reading process can be distinguished different levels. These vary from the recognition of the graphic characteristics to the text interpretation. The primary

assumption is that information processing consists of parallel steps that interact simultaneously and continuously, rather than hierarchical steps that occur one after another (Carrell et al., 1988). Stanovich (1980) introduced the notion of activation in his interactive-compensatory model. It assumes that the interpretation of a text is synthesized from information provided simultaneously from various knowledge sources, and that there is a compensation between them, which differentiates strong readers from weaker ones. The interaction among the knowledge sources also refers to the schema theory, which presumes the text does not carry meaning by itself, but only provides indications for the reader to construct meaning from their own patterns of knowledge (Adams & Collins, 1979; An, 2013). Effective understanding of a text is therefore based on the ability to connect the elements of the text to prior knowledge (An, 2013).

The construction of meaning is achieved in the interactions between text, reader and context, as noted in the contemporary reading model (Giasson, 1996; Irwin, 1991). The understanding of text varies according to the degree of association among these three components. The more intertwined they are, the better the understanding of text (Giasson, 1996). Irwin (1991) classifies reading processes into five categories: 1) microprocesses help readers understand the information in a sentence; 2) integrative processes establish connections between sentences; 3) macroprocesses relate to overall understanding and coherent links within the text; 4) elaborative processes involve making inferences, and 5) metacognitive processes manage comprehension and allow readers to adjust to the reading situation (Giasson, 1996). On the other hand, Kintsch (1988) and Van Dijk and Kintsch (1983) distinguish between two levels of reading comprehension in the construction-integration model: global comprehension (the macrostructure) involves the entire text and local comprehension (the microstructure) relates to the reading of each sentence or paragraph.

This constructive and interactive view of reading translates into two reading purposes on the PIRLS test: 1) reading for literary experience and 2) reading to acquire and use information. Through narrative passages, reading for literary experience (purpose 1) asks students to explore unreal situations by bringing to the text their own experiences and feelings, as

well as appreciation of language and their knowledge, to interpret and create meaning within the text (Mullis et al., 2009). These ideas reflect fundamental characteristics of the construction-integration model of reading (Kinstch, 1988) and the schema theory (Anderson & Pearson, 1984). By reading to acquire and use information (purpose 2), students engage with the facts of informational texts to understand how the events come about (Mullis et al., 2009). Differences in the organization and structure of informational texts require readers to use a variety of cognitive and metacognitive strategies to respond to the intended tasks, which has been emphasized in interactive reading models (Rumelhart, 1978).

Meaning is constructed in the PIRLS test through four reading processes: 1) examine and evaluate content, language and textual elements, 2) make straightforward inferences, 3) focus on and retrieve explicitly stated information and 4) interpret and integrate ideas and information (see Table 1).

To focus on and retrieve explicitly stated information (P3), readers use different ways to answer the question by identifying explicitly stated information in the text. This requires understanding sentences without resorting to interpretation or inference (Labrecque et al., 2012). On the other hand, making straightforward inferences (P2) requires that students go beyond the surface of the text and fill in the gaps by establishing relationships between various types of information (Labrecque et al., 2012; Mullis et al., 2009). These two processes are part of a local comprehension local understanding.

Conversely, to interpret and integrate information in the text (P4), readers make implicit connections from their own perspective using prior knowledge, which reflects the schema theory (Anderson & Pearson, 1984). The level of integration and interpretation of information varies according to the experiences and knowledge harnessed for the tasks (Labrecque et al., 2012; Mullis et al., 2009). Lastly, in order to examine and evaluate content, language, and textual elements (P1), readers must step back from the text to look critically at the content, language, or textual elements by considering genre, structure, or linguistic conventions (Labrecque et al., 2012). This process corresponds to the global comprehension of the text.

Table 1
Description of reading processes and distribution of items

Process		Definition	No. of items
P1	Examine and evaluate content, language and textual elements	Take a critical look at the content, language used or textual elements and reflect on the clarity of how meaning is expressed by drawing on one's knowledge of the genre, structure, or linguistic conventions	4
P2	Make straightforward inferences	Fill in the gaps in meaning by inferring information from the text	11
P3	Focus on and retrieve explicitly stated information	Understand and locate the information stated explicitly, and make the link with the question	7
P4	Interpret and integrate ideas and information	Acquire a deeper understanding of the text by combining prior knowledge and information presented in the text	13

Cognitive diagnostic approach

CDA was developed during the 1980s with two main components: 1) content analysis of items to identify underlying cognitive attributes and 2) psychometric models representing the relationships between these items and attributes (Lee & Sawaki, 2009; Yang & Embretson, 2007). Attributes refer to skills, knowledge and cognitive strategies that the student uses to correctly answer the items (Buck & Tatsuoka, 1998; Lee & Sawaki, 2009; Leighton & Gierl, 2007). In practice, the diagnosis in this approach are determined in two ways. The first is to analyze data from large-scale tests, which were not designed for diagnostic purposes, using a cognitive model to extract detailed information about students' skill mastery. The second is to design a test for diagnostic purposes and then analyze the test results to obtain diagnostic information (DiBello, Roussos & Stout, 2007).

Our research is based on the first approach and consists of four steps: 1) attribute identification, 2) Q-matrix development, 3) data modelling and 4) diagnostic feedback:

1. Attribute identification is often performed by a panel of experts by analyzing test specifications, item content, underlying theoretical models, and empirical research findings (Lee & Sawaki, 2009; Leighton & Gierl, 2007);
2. A Q-matrix is then created to represent the relationships between the attributes identified and the items. This matrix is often in the form of an array with binary numbers (1 and 0) to determine whether an attribute is required to correctly answer an item. The Q-matrix can be constructed by combining a content analysis of the items and an analysis of empirical data from a small group of candidates (Loye & Lambert-Chan, 2016; Tjoe & de la Torre, 2014) or analysis of candidates' think-aloud protocol (Jang, 2005; Li & Suen, 2013);
3. The developed Q-matrix is integrated into data modelling with the diagnostic classification models (DCMs). These models are based on the premise that student performance depends on mastery or non-mastery of a set of attributes that cannot be directly observed (Buck & Tatsuoka, 1998);
4. Data modelling provides results on item parameters that evaluate the diagnostic quality of test items, the quality of the Q-matrix, and subject parameters that provide information regarding students' attribute mastery profiles (Loye, 2010). Student profiles are classified in binary form (0 = non-mastery and 1 = mastery) based on a cut-off point of 0.5. This cut-off point has been suggested in a biased manner in the research of Li (2011), Lee and Sawaki (2009) and de la Torre (2009). Other research (Jang, 2005, 2009) classifies profiles in a multicategory form: non-mastery if the probability of mastery (p) is 0 to 0.4; undetermined profile (p = 0.41 to 0.6); and mastery (p = 0.61 to 1).

Literature review of research conducted with DCMs

The first CDA studies on reading were conducted using rule-space model (Buck, Tatsuoka & Kostin, 1997; Kasai, 1997; Scott, 1998). Buck and his colleagues (1997) analyzed data from 5,000 Japanese students who took the Test of English for International Communication (TOEIC). The attributes identified are based on the taxonomy of sub-skills proposed by

Grabe (1991) and on empirical studies (Freedle & Kostin, 1993). With seven attributes identified, the authors were able to classify 91% of the candidates into their skill mastery profiles and provide probabilities of mastery for each skill. These scores were then analyzed with multiple regression and the results suggest that attributes can explain 97% of the variation in candidate performance. Thus, the rule-space model can explain students' performance of complex tasks like reading and can provide diagnostic information (Buck et al., 1997). However, the limitations of this study lie in the subjectivity of the attribute selection criteria. Moreover, data analysis is performed with only one form of the test but could have been done with other formats in order to compare the results obtained (Buck et al., 1997).

Jang (2009) modelled the data of 2,703 TOEFL iBT candidates with the fusion model (Hartz, 2002). By analyzing learners' think-aloud protocols, the experts identified nine skills, namely: 1) inferring the meaning of a word or sentence, 2) determining the meaning of a word using prior knowledge, 3) understanding relationships between parts of the text using logical connectors, 4) identifying explicitly stated information, 5) understanding implicitly stated information, 6) making inferences, 7) formulating negation 8) summarizing main ideas, and 9) recognizing contradictory ideas or arguments. Skills 8, 1, 2 are most mastered by learners, whereas skills 7 and 9 are least mastered. The interesting thing about this research is that students' performance on the test was assessed before and after they took preparatory courses. The results show that their probability of mastering the skills was increased by 12% after the course and that approximately 85% of students can improve their skills performance.

This model was also used to analyze data from the MELAB test in Li (2011) and Li and Suen (2013). The attributes identified are based primarily on research by Gao (2006) and Jang (2005). With the think-aloud protocols, the authors initially identified six attributes, but eventually reduced them to four, given the insufficient number of items per attribute: 1) vocabulary, 2) syntax, 3) extraction of explicit information, and 4) understanding of implicit information. The results suggest that the attributes are mastered by 55% of students overall. However, the diagnostic quality is still low for some items, as the test was not specifically designed for diagnostic purposes (Li, 20; Li & Suen, 2013).

Lee and Sawaki (2009, 2011) conducted a study with the TOEFL iBT to diagnose reading and listening performance with four attributes: 1) understanding vocabulary, 2) understanding specific information, 3) connecting information and synthesizing, and 4) organizing information. The data were analyzed using the fusion model, the general diagnostic model (GDM) and the latent class model (LCM), which allow to compare the candidates' mastery profiles. Von Davier (2008) also applied the GDM model to the TOEFL iBT by analyzing two test formats, that deal with both dichotomous and polychotomous data. With the four skills identified, the results demonstrated the applicability of GDM to large-scale language testing.

In the context of large-scale international tests, a number of CDA research were conducted with mathematics tests, such as the TIMSS. Lee, Park and Taylan (2011) modelled data from 823 Grade 4 students who completed booklets 4 and 5 of TIMSS 2007 with the deterministic inputs, noisy and gate model (DINA). A total of 15 attributes were identified for the test, which covers three domains: 1) whole numbers, 2) geometry and 3) data display and interpretation of results. The same test was used in the research of Evran (2019), Arican and Sen (2015), Terzi and Sen (2019), Toker and Green (2012), Wafa (2019) and Wafa, Hussaini and Pazhman (2020). The research results with an international test such as TIMSS thus support the premise that the CDA provides more detailed information about students' mastery of attributes than traditional methods, which can be directly used to improve classroom instruction (Lee, Park and Taylan, 2011).

DINA and G-DINA models

The deterministic inputs, noisy and gate model (DINA) is a non-compensatory model which assumes that the participant must master all the attributes necessary to answer items correctly. For each item, this model divides candidates into two latent classes: 1) those who master all the attributes required for one item ($\xi_{ij} = 1$) and 2) and those who don't master them ($\xi_{ij} = 0$) (Cui, Gierl and Chang, 2012; de la Torre and Douglas, 2008; Junker and Sijtsma, 2001). The model takes into account that the candidate may give the wrong answer, even if they master all the required attributes (Loye, 2010). It therefore considers two parameters: 1) the guessing parameter (g_i), which refers to the probability that an individual can answer an item correctly, even if they do not master all the necessary

attributes, and 2) the slipping parameter (s_i), which represents the probability that an individual can give the wrong answer, even if they master all the required attributes. Ideally, these parameters should be small to show a high diagnostic quality for the item. The relationships between these parameters are represented as follows:

$$P(X_{ij} = 1 \mid \xi_{ij}, s_j, g_j) = (1 - s_j)^{\xi_{ji}} \; g_j^{1-\xi_{ji}}$$

Thus, for the group that masters all the attributes, the probability of answering an item correctly is equal to $1-s_i$, while for those who do not master them, this probability is equal to g_i. Table 2 summarizes these probabilities according to the two latent groups.

Table 2
Response probabilities in the DINA model
(adapted from Rupp, Templin and Henson, 2010)

	$X_{ij} = 1$ **(Correct response)**	$X_{ij} = 0$ **(Wrong response)**
$\xi_{ij} = 1$ **Mastery of all attributes**	$1\text{-}s_i$	s_i
$\xi_{ij} = 0$ **Non-mastery of all attributes**	g_i	$1\text{-}g_i$

Unlike DINA, generalized DINA (G-DINA) does not take into account the restricted conjunctive or disjunctive relationship of attributes to correctly answer an item (de la Torre, 2011; Ravand, Barati & Widhiarso, 2013). Thus, instead of separating participants into two latent classes for each item, G-DINA divides them into $2^{K^*_j}$ latent groups, where K^*_j is the number of attributes required for item j. Each group represents an attribute vector reduced to α^*_{ij}, which has its own probability of success (de la Torre & Douglas, 2008). G-DINA assumes that, even if candidates cannot master all the necessary attributes ($\xi_{ij} = 0$), the probability of getting a correct answer may vary.

We selected these models for the data analysis for three reasons. First, they are not yet widely used in the language field, whereas they have been successfully applied to mathematical data or simulation study (Cui, Gierl &

Chang, 2012; de la Torre & Douglas, 2008; de la Torre, 2011). Second, these models are the simplest and therefore the most restrictive and interpretable of the CDAs that can manage dichotomous data (de la Torre & Douglas, 2008). According to DiBello, Roussos and Stout (2007), when selecting a CDA, the feasibility and the parsimony must be considered. These aspects relate to the importance of keeping the models as simple as possible in terms of parameters and an appropriate fit of the data to achieve the diagnostic purposes. Third, these models can compensate for each other, given that G-DINA can overcome a major limitation of the DINA model since it can differentiate participants with different levels of probability of answering correctly, even if they do not master all the required attributes. For example, for an item that requires three attributes, the participant who masters two has a higher probability of success than the participant who masters only one attribute (Ravand, Barati & Widhiarso, 2013).

Methodology

This study investigates the feasibility of diagnostic modelling of the data from PIRLS 2011. Specifically, we assess: 1) the fit of the DINA and G-DINA models to the data with the two Q-matrices, 2) the diagnostic quality of the items and 3) the skill mastery profiles of the students.

Database

The test is in English and consists of 35 items, divided into two parts that correspond to two reading purposes. The first part is a literary passage with 16 items, while the second part is an informational passage with 19 items. There are 15 multiple-choice questions (MCQs) that are worth one point each with 4 answer options. These questions are used to evaluate comprehension processes that do not require judgment or complex interpretation. In addition, 20 items are constructed-response questions worth either two (19 items) or three points (1 item). These questions assess the process of interpretation (P4), which requires the use of students' prior knowledge and experience (Labrecque et al., 2012).

In Canada, 23,206 students participated in the PIRLS 2011 test, divided into 13 booklets. Of these, 16,500 students took the test in English, while approximately 6,500 students took the test in French (Labrecque et

al., 2012). The database selected for this research contains responses from 4,762 students who completed booklet 13, 49.3% of whom were female and 50.7% male. Student responses were coded dichotomously. Answers to MCQs were coded as 1 (correct answer) or 0 (incorrect answer). For constructed-response questions, 0 and 1 point responses were coded as 0, while 2 and 3 point responses were coded as 1. Missing data or incomplete responses were considered wrong answers and were coded as 0. This way of coding data has been used in research with the TIMSS by Lee, Park and Taylan (2011) and Evran (2019).

Participants

To identify the underlying attributes of the test, a panel of experts comprising three members was formed according to three criteria: 1) good knowledge of languages, 2) experience in language teaching, and 3) experience in analyzing data from large-scale diagnostic language tests. Expert #1 had experience in developing the Q-matrix for the Programme for International Student Assessment (PISA) test, while the second had experience in identifying test specifications for the French test for immigrants in Quebec. Expert#3 was the co-researcher herself, who had a background in language teaching and a good knowledge of CDA.

Procedure for developing Q-matrices

Two Q-matrices were developed for our research. It should be noted that the PIRLS test was designed according to a well-developed framework with underlying cognitive reading models. This framework identified four reading processes that students must use to answer the questions, which constitute our Q1-matrix (see Table 3). The idea is to verify whether these processes could be used as attributes for diagnostic modelling. However, since each item on the PIRLS test corresponds to a single reading process, it risks providing a more global diagnosis. Thus, we are interested in whether it is possible to decompose each process into multiple cognitive reading strategies with the expert panel. Hence, the idea of developing a Q2-matrix (see Table 5) and comparing these two matrices to determine which one works better with the data. The Q1-matrix was developed by the co-researcher from four comprehension processes identified in the PIRLS 2011 framework. Because each item assesses one comprehension process, the 35 items were one attribute-items. Table 3 summarizes the descriptions of processes and the distribution of items by reading process.

Table 3
Q1-matrix developed from the PIRLS 2011 test framework

Extract	Item	P1	P2	P3	P4
		Examine and evaluate content, language and textual elements	**Make straight-forward inferences**	**Focus on and retrieve explicitly stated information**	**Interpret and integrate ideas and information**
Passage 1 Literary text	**1**	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	1	0	0
	6	0	0	1	0
	7	0	0	1	0
	8	0	1	0	0
	9	0	1	0	0
	10	0	1	0	0
	11	0	1	0	0
	12	0	0	0	1
	13	1	0	0	0
	14	0	0	0	1
	15	0	0	0	1
Passage 2 Informational text	**16**	1	0	0	0
	17	0	0	1	0
	18	0	1	0	0
	19	0	0	1	0
	20	0	0	0	1
	21	0	1	0	0
	22	0	0	1	0
	23	0	1	0	0
	24	0	0	0	1
	25	0	0	0	1
	26	0	0	0	1
	27	0	1	0	0
	28	0	0	0	1
	29	0	0	1	0
	30	1	0	0	0
	31	0	0	0	1
	32	0	0	0	1
	33	0	0	0	1
	34	0	0	0	1
	35	0	1	0	0
	Total	**4**	**11**	**7**	**13**

In this Q1-matrix (see Table 3), we note an unequal distribution of processes among items and passages. For example, P1 (Process 1) has only four items (11.43%): two in passage 1 and two in passage 2. But P4 has the largest number of items: four in passage 1 and nine in passage 2, for a total of 13 items (37.14%). P2 has a total of 11 items (31.43%), including six items in passage 1 and five in passage 2. Lastly, seven items (20%) are linked to P3, including three items in passage 1 and four in passage 2. For the Q2-matrix, the experts identified a list of underlying attributes to answer the items correctly. The co-researcher provided each expert with 1) the content of the two passages, the items, and the answer key, 2) the test development framework, 3) information about the item parameters and 4) instructions and expectations for the required tasks. Initially, Expert #1 worked with Expert #3 to develop a list of attributes necessary for the test. By analyzing the content of the passages, questions, and answer key for booklet 13, the experts came up with five attributes needed for the test (see Table 4), with some questions requiring more than one comprehension process to answer the item.

To illustrate, P4, or "Interpret and integrate ideas and information," asks students to identify the overall message or theme of a passage, highlight commonalities and differences in information, and interpret possible real-life applications of information (Labrecque et al., 2012). These tasks require a global understanding and interpretation of ideas in their own words. We therefore separated this process into two attributes: global comprehension (A2) and interpretation (A3).

As for P1, "Examine and evaluate content, language, and textual elements", students must compare the connotation of a word to their own understanding, or to information from other sources, and reflect on the clarity of the expressed meaning, using their own knowledge (Labrecque et al., 2012). This process refers to an overall understanding of the text and also to students' ability to reformulate ideas in their own words, which requires a mastery of vocabulary and syntax.

By analyzing all the reading processes stated in the framework, the two experts ultimately identified five attributes: A1 Identifying explicit information, A2 Global comprehension, A3 Interpretation, A4 Making straightforward inferences and A5 Vocabulary and syntax. Table 4 presents a more detailed definition of these five attributes.

Table 4
Detailed attribute definitions

Attribute		Definition
A1	Identifying explicit information	Locating and recognizing explicit information in the text to answer questions
A2	Global comprehension	Forming an overall understanding of a paragraph or of the entire text
A3	Interpretation	Clarifying the meaning of complex ideas or configurations and interpreting relationships
A4	Making inferences	Understanding information not explicitly stated by making inferences and predictions
A5	Vocabulary and syntax	Expressing ideas in grammatically correct and clear written English

Then, the experts individually identified the necessary attributes for each item. Because Expert #2 did not participate in identifying the list of attributes, we asked him to add more attributes, if he felt it was necessary. Experts could choose more than one attribute per item if needed. Here, they were to rank these attributes in order of importance. Table 5 identifies the experts' attributes for each item.

When the results were submitted, no attributes were added by Expert #2. Fleiss' (1971) kappa statistic was calculated to measure the degree of inter-judge agreement using AgreeStat 2015 software. Table 6 shows the percentages of items by inter-judge agreement. Only items approved by at least two of the three experts (Fleiss kappa ≥ 0.6) were accepted. Items with greater agreement were logically one-attribute items. For items with low and medium agreement, we reviewed the experts' comments to make necessary adjustments.

The selected attributes for each item were then compiled to form an initial Q2-matrix. This matrix was examined and refined by our experts to arrive at a final Q2-matrix, which was retained for modelling purposes. Although strict guidelines are not proposed, Hartz (2002) suggests that each attribute should be measured by at least three items and be broadly defined. Thus, attributes that do not meet this criterion are combined with either a similar attribute or eliminated from the final Q2-matrix. Lastly, the final Q2- matrix (see Table 7) contains 9 one-attribute items, 21

Table 5
Initial Q2-matrix developed by the experts

	Item	Experts			Proposed attributes[1]
		1	**2**	**3**	
Passage 1 Literacy text	1	1; 2	2	2; 4	2; 4
	2	4; 5	3; 5	3; 5	3; 5
	3	1	1	1	1
	4	3; 4	3; 4	3; 4	3; 4
	5	1; 4; 5	4; 5	1; 4; 5	1; 4; 5
	6	1	1	1	1
	7	1	1	1	1
	8	1	4	4	4
	9	4; 5	1; 3; 5	3; 5	3; 5
	10	1; 4	3; 4	3; 4	3; 4
	11	3; 4	1; 3; 4	3; 4	3; 4
	12	3; 4	3; 4	1; 3; 4	3; 4
	13	2; 4	2; 3; 4	2; 4	2; 4
	14	3; 5	2; 3; 5	2; 3; 5	2; 3; 5
	15	1; 2; 3; 5	1; 2; 3; 4	1; 2; 3; 4; 5	1; 2; 3; 4; 5
Passage 2 Informational text	16	2; 5	2; 3	2; 3; 5	2; 3; 5
	17	1	1	1	1
	18	1; 3	1; 3	1; 3	1; 3
	19	1	1	1	1
	20	1; 3	1; 2; 3	1; 3	1; 2; 3
	21	4	3; 4	3; 4	3; 4
	22	1	1	1	1
	23	1; 4	4	3; 4	3; 4
	24	3; 5	3; 5	3; 5	3; 5
	25	3; 5	3; 5	3; 5	3; 5
	26	3; 5	3; 5	3; 5	3; 5
	27	4	3	3; 4	3; 4
	28	3; 5	2; 3; 5	3; 5	3; 5
	29	1	1	1	1
	30	2	2; 5	2; 5	2; 5
	31	1; 3	1; 3	1; 3	1; 3
	32	1; 3	3	1; 3	1; 3
	33	1; 3	1	1; 3	1; 3
	34	1; 3	1; 3	1; 3	1; 3
	35	1; 3	1	1	1

1.Attributes selected in the discussion with the experts.

Table 6
Percentages of items according to the experts' agreement rate

Agreement rate	**Fleiss kappa**	**% of items**
Weak	0 à 0,4	17,2%
Medium	0,41 à 0,6	11,4%
Strong	0,61 à 0,8	31,4%
Perfect	0,81 à 1	40,0%

two-attribute items, 4 three-attribute items, and only 1 five-attribute item. The one- and two-attribute items are worth 1 point, whereas the three- and five-attribute items are worth 2 and 3 points, respectively.

Table 7
Proposed final Q2-matrix for PIRLS 2011 Test

	Item	A1 Identifying explicit information	A2 Global comprehension	A3 Interpretation	A4 Making inferences	A5 Vocabulary and syntax
Passage 1 Literacy text	**1**	0	1	0	1	0
	2	0	0	1	0	0
	3	1	0	1	0	0
	4	0	0	1	1	0
	5	1	0	0	1	1
	6	1	0	1	0	0
	7	1	0	1	0	0
	8	0	0	0	1	0
	9	0	0	1	0	1
	10	0	0	1	1	0
	11	0	0	1	1	0
	12	0	0	1	1	0
	13	0	1	0	1	0
	14	0	1	1	0	1
	15	1	1	1	1	1
Passage 2 Informational text	**16**	0	1	1	0	1
	17	1	0	0	0	0
	18	1	0	1	0	0
	19	1	0	0	0	0
	20	1	1	1	0	0
	21	0	0	1	1	0
	22	1	0	0	0	0
	23	0	0	1	1	0
	24	0	0	1	0	1
	25	0	0	1	0	1
	26	0	0	1	0	1
	27	0	0	1	1	0
	28	0	0	1	0	0
	29	1	0	0	0	0
	30	0	1	0	0	1
	31	1	0	1	0	0
	32	1	0	1	0	0
	33	1	0	1	0	0
	34	1	0	1	0	0
	35	1	0	0	0	0

Analysis

The dichotomous database and the Q1 and Q2 matrices were modelled with DINA and G-DINA using OxEdit software, allowing assessment of model relative and absolute fit to the data, estimate item parameters and identify students' skill mastery profiles.

Evaluation of model fit to the data

Evaluation of model fit to the data enables verification of the fundamental consistency between the estimated model (predicted data) and the observed data to suggest improvements to the model (DiBello, Roussos & Stout, 2007; Sinharay, 2004). In evaluating the fit of DCMs, we can distinguish between relative fit and absolute fit. Thus, evaluating the relative fit of DCMs refers to the process of selecting the most appropriate model from competing models (Chen, de la Torre & Zhang, 2013). In this study, the following three statistics are used for assessing the relative fit of DINA and G-DINA:

1) -2 log-likelihood (-2LL): $-2LL = 2\ln(ML)$

2) Akaike information criterion (AIC): $-2LL + 2P$

3) Bayesian information criterion (BIC): $-2LL + P\ln(N)$,

where ML is the maximum likelihood of the item parameters; P is the number of model parameters; L is the total number of attribute patterns and N is the sample size. For each statistic, the model with the smallest value will be preferred over competing models (Chen, de la Torre & Zhang, 2013).

The evaluating of the DCMs absolute fit determines whether the models fit the data adequately. Thus, three statistics are used: 1) the residual between the predicted and observed proportion of correct items, 2) the residual between the predicted and observed Fisher transformation for each pair of items, and 3) the residual between the observed and predicted log-odds ratio of each pair of items (Chen, de la Torre & Zhang, 2013). With these three statistics, a large number of attribute patterns is sampled from the posterior distribution of attributes. The generalized attribute patterns and estimated parameters can be used to generate predicted item responses. The difference between the observed and predicted responses should be 0 if the model fits the data adequately.

To use these three statistics, we need to calculate their standard errors (SE), which allows us to derive Z-scores from these three statistics to check if the residuals are statistically different from 0. Rejecting any Z-score means that the model does not fit an item or pair of items adequately (Chen, de la Torre & Zhang, 2013). We must rely on at least two of the three indices that are statistically different from 0 to show that the selected model fits the data adequately. Thus, this step of assessing the absolute fit allows us to detect errors of overspecification or underspecification of the attributes in the Q-matrix.

Evaluation of the diagnostic quality of the items

By estimating the guessing and slipping parameters, we can evaluate the diagnostic quality of the items, which is determined by 1-*g*-*s*. Thresholds for interpreting parameters are often biased and vary from author to author. For example, according to de la Torre (2009), the diagnostic quality of items can be classified into three categories: 0-0.1 = high quality; 0.1-0.2 = medium quality; and 0.2-0.3 = low quality. In contrast, according to Ma, Iaconangelo and de la Torre (2016), items are classified as high quality if these parameters are between 0 and 0.15; medium quality if they are between 0.15 and 0.25; and low quality if they are between 0.25 and 0.35. And finally, according to Ravand, Barati, and Widhiarso (2013), items are considered high quality if these parameters are below 0.5 and low quality if they are above 0.5.

Results

Evaluation of the relative and absolute fit of the models to the data

Relative fit

Table 8 presents the results of the relative fit of models to the data with the -2LL, AIC and BIC indices. The model with the lowest values is the one that best fits the data. Thus, the results suggest that the G-DINA fits the data better than the DINA with the final Q2-matrix, compared to the Q1-matrix. Specifically, with DINA and the Q1-matrix, the indices are slightly lower than those with G-DINA. However, with the final Q2-matrix, the G-DINA fits better than the DINA, given the lower values of these

Table 8
Models relative fit to data with Q1 and Q2- matrices

DCM	Q-Matrix	-2LL	AIC	BIC
DINA	Q1	171273.2021	171443.2021	171993.0181
	Q2	170382.2495	170584.2495	171237.5603
G-D INA	Q1	171274.9403	171444.9403	171994.7562
	Q2	163575.4671	163969.4671	165243.7464

Note. -2LL = -2 log-likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion.

statistics. In both models, the data fit better with the final Q2-matrix than with the Q1-matrix. We therefore conclude that the G-DINA model is the best fit to the data with the final Q2-matrix.

Absolute fit

For the absolute fit assessment, the correct proportion (prop.), transformation correlation (Z [Corr]) and log-odds ratio (Log [OR]) statistics were used (Chen, de la Torre and Zhang, 2013) (see Table 9). These values must be close to 0 for all items to show that the model fits the data. The maximum Z-score values for these three statistics were also derived. In addition, the rejection thresholds of these Z-scores were used to decide whether the models fit the data adequately or not. In principle, these values should be below the critical values to show that the selected model fits the data adequately. If not, the fit of the selected model is rejected (Ma & Meng, 2014).

Correct proportion

For the correct proportion (see Table 9), the maximum Z-score values are more or less similar to the two matrices in both the DINA and G-DINA models. The maximum Z-score values are lower with the G-DINA model and lower with the final Q2-matrix. Comparing the critical values of the Z-scores with the Bonferroni correction, these values are all lower than the critical values. We therefore conclude that both models fit the data adequately with all items and with both Q-matrices.

Table 9
Models absolute fit to data with Q1- and Q2-matrices

Matrix		Prop.		Z (Corr)		Log (OR)	
		DINA	G-DINA	DINA	G-DINA	DINA	G-DINA
Max	Q1	0.0125	0.0076	1.2399	0.9744	12.2611	11.7888
	Q2	0.0089	0.0062	0.9880	0.3658	11.1028	8.4344

Note. Max = maximum value of Z-scores; Prop = correct proportion; Z (Corr) = transformation correlation; Log (OR) = log-odds ratio. Critical Z-score value (Zc) = 3.467; 3.649; 4.044 for $\alpha = 0.1$; 0.05; 0.01, respectively (with Bonferroni correction).

Transformed correlation

With the Fisher transformed correlation, the maximum values for DINA and G-DINA with the Q1-matrix are close. The difference is larger between DINA and G-DINA in the final Q2-matrix. According to the correct proportion (prop.), the G-DINA model fits the data better and it fits better with the final Q2-matrix. The comparison with the critical values of the Z-scores confirms that both models fit the data adequately for all items and with both Q-matrices.

Log-odds ratio

As regards the Z-scores of the log-odds ratio, the values obtained differ significantly from the correct proportion and the transformed correlation, although we observe the same tendency, i.e., the values are lower with the G-DINA model and lower with the final Q2-matrix. However, these values are all higher than the critical values of the Z-scores. We therefore conclude that the DINA and G-DINA models do not fit the data adequately with all the items.

Comparing these three statistics, we find that the values are close between the correct proportion and the transformed correlation in both models and with both Q-matrices. This leads to the same decision to not reject the hypothesis null and to conclude that the models fit the data adequately with all the items. This decision suggests the sensitivity of these two statistics in assessing the absolute fit to the data is almost similar. On the other hand, the values of the log-odds ratio differ considerably from the two previous statistics and lead us to reject the the models don't fit

the data adequately, so the models don't fit the data adequately with all items. This decision leads us to question the reliability and sensitivity of this statistic in assessing the model absolute fit the data.

To sum up, the results show that the G-DINA fits the data better than the DINA, and better with the final Q2-matrix than the Q1-matrix. These models fit the data adequately with both Q-matrices according to the correct proportion and transformed correlation statistics. However, the results of the log-odds ratio tell us otherwise. We therefore rely on the results of the correct proportion and the transformed correlation, as they lead to the same decision. Based on these results, we examine the item parameters with the DINA model and the students' skill mastery profiles obtained with the final Q2-matrix and the G-DINA.

Items parameters estimates

Guessing parameter

The average of guessing parameter is 0.36443 (see Table 10), i.e., a student has a 36.43% of probability of answering the questions correctly, even though they have not mastered all the required attributes. According to the criteria defined by de la Torre (2009), as guessing parameter, there are six high quality items, three medium quality items and seven low quality items. In total, 19 items are problematic. Most of the high and medium quality items are found in Part 2 of the test. Part 1 contains only four medium and low-quality items. According to Ravand, Barati, and Widhiarso's (2013) criteria, there are 23 high quality items and 12 low quality items in terms of guessing parameter.

Slipping parameter

The average of slipping parameter is 0.2198 (see Table 10). In other words, on average, students have a 21.98% of chance of answering incorrectly, even though they have mastered all the required attributes. According to the criteria defined by de la Torre (2009), in terms of slipping parameter there are 13 high quality items, four medium quality items, and seven low quality items. Finally, 11 items are problematic as slipping parameter. However, the thresholds suggested by Ravand, Barati, and Widhiarso (2013), there are only three low quality items and 32 high quality items.

Table 10
Item parameter estimates with final Q2-matrix and with DINA

	Item	*g*	SE	*s*	SE	*s*+*g*
Passage 1 Literacy text	**1**	0.6775	0.0110	**0.0466**	0.0049	0.7241
	2	0.7340	0.0087	**0.0612**	0.0058	0.7952
	3	0.5769	0.0125	**0.0865**	0.0053	0.6634
	4	**0.3088**	0.0105	**0.3246**	0.0100	0.6334
	5	**0.2295**	0.0088	**0.2741**	0.0114	**0.5036**
	6	0.5901	0.0124	**0.0898**	0.0053	0.6799
	7	**0.2338**	0.0113	**0.3441**	0.0088	**0.5779**
	8	0.5538	0.0129	**0.0798**	0.0058	0.6336
	9	0.7147	0.0088	**0.0525**	0.0053	0.7672
	10	0.7160	0.0097	**0.0083**	0.0022	0.7243
	11	0.5624	0.0110	**0.0693**	0.0058	0.6317
	12	**0.4194**	0.0111	**0.0981**	0.0068	**0.5175**
	13	0.6917	0.0107	**0.0152**	0.0031	0.7069
	14	**0.4405**	0.0097	**0.1383**	0.0088	**0.5788**
	15	**0.1958**	0.0077	**0.3179**	0.0121	**0.5137**
Passage 2 Informational text	**16**	**0.2708**	0.0088	**0.3898**	0.0121	0.6606
	17	0.6136	0.0123	**0.1027**	0.0057	0.7163
	18	**0.3947**	0.0109	**0.3441**	0.0092	0.7388
	19	0.5280	0.0127	**0.1384**	0.0065	0.6664
	20	**0.0862**	0.0060	0.5939	0.0109	0.6801
	21	**0.3806**	0.0109	**0.2656**	0.0095	0.6462
	22	0.5345	0.0127	**0.1614**	0.0069	0.6959
	23	**0.3923**	0.0110	**0.2013**	0.0088	**0.5936**
	24	**0.0000**	0.0066	0.5207	0.0117	**0.5207**
	25	**0.2618**	0.0090	**0.2390**	0.0105	**0.5008**
	26	**0.0265**	0.0035	**0.4474**	0.0118	**0.4739**
	27	**0.4428**	0.0111	**0.3008**	0.0098	0.7436
	28	**0.2642**	0.0091	**0.2885**	0.0109	**0.5527**
	29	**0.2764**	0.0117	**0.2879**	0.0085	**0.5643**
	30	**0.0425**	0.0046	**0.7207**	0.0107	0.7632
	31	**0.0117**	0.0035	**0.0189**	0.0032	**0.0306**
	32	**0.1202**	0.0074	**0.0606**	0.0048	**0.1808**
	33	**0.1602**	0.0083	**0.0777**	0.0053	**0.2379**
	34	**0.0339**	0.0042	**0.1999**	0.0078	**0.2338**
	35	**0.2643**	0.0115	**0.3271**	0.0088	**0.5914**
	Moyenne	**0.3643**		**0.2198**		**0.5841**

Note. *g* = pseudo-likelihood; SE = standard error; *s* = forgetting.

Guessing and slipping parameters

The sum of the averages of the two parameters is 0.5841, indicating that the diagnostic quality of the items is 0.4259. According to the thresholds defined by de la Torre (2009), there are only two high quality items ($g+s$ = 0 to 0.2); there are two medium quality items ($g+s$ = 0.2 and 0.4); whereas 12 are of low quality items ($g+s$ = 0.4 and 0.6). In the end, 19 items were considered problematic ($g+s$ > 0.6). Most of them appear in Part 1 of the test. But according to the thresholds suggested by Ravand, Barati, and Widhiarso (2013), there are 16 high quality items ($g+s$ < 0.6) and 19 low quality items ($g+s$ > 0.6). Figure 1 shows the estimated item parameters and the diagnostic quality of the items. The higher the line, the better the diagnostic quality of the items.

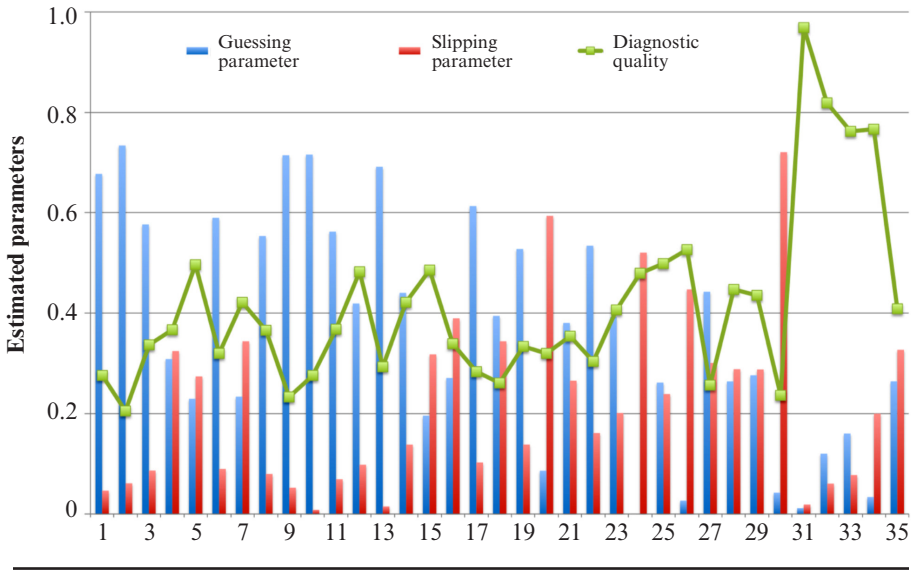

Figure 1. Items parameters estimates and their diagnostic quality

Students' skill mastery profiles

Table 11 shows the 32 profiles with the corresponding percentage of students per profile. The most popular profile is 11110 (25.24%), meaning that 25.24% of students master the first four attributes, but not the fifth. Next is the profile of students who master none of the five attributes, 00000 (18.78%). The profile of those who master all attributes (11111)

ranks third (16.13%). The 10011 profile also occurs among our participants at 9.11%. According to this profile, the student has mastered the attributes A1 *Identifying implicit information*, A4 *Making inferences*, and A5 *Vocabulary and syntax*, but not the attributes A2 *Global comprehension* and A3 *Interpretation*. Profile 00011 follows closely behind at 8.89%. No participant belongs to one of the following four profiles: 10000, 11000, 01001 and 11010. We will provide more details in the Discussion section.

Table 11
Skill mastery profiles and percentages of participants

Profile	% of students	Profile	% of students
00000	18.78	11100	2.06
10000	0	11010	0
01000	0.35	11001	0.01
00100	4.12	10110	1.97
00010	0.40	10101	0.47
00001	2.13	10011	9.11
11000	0	01110	0.55
10100	2.3	01101	0.17
10010	0.15	01011	0.06
10001	0.19	00111	0.03
01100	2.19	11110	25.24
01010	0.12	11101	0.24
01001	0	11011	3.18
00110	00.48	10111	0.32
00101	0.37	01111	0.01
00011	8.89	11111	16.13

A4 *Making inferences* is the most mastered attribute (66.62%), followed by A1 *Identifying implicit information* (61.31%). A3 *Interpretation* is ranked third (56.64%) in the probability of mastery. A2 *Global comprehension* is mastered by 50.31% and A5 V*ocabulary and Syntax* is the least mastered (41.31%). Figure 2 presents the probabilities of mastering the attributes for all the students.

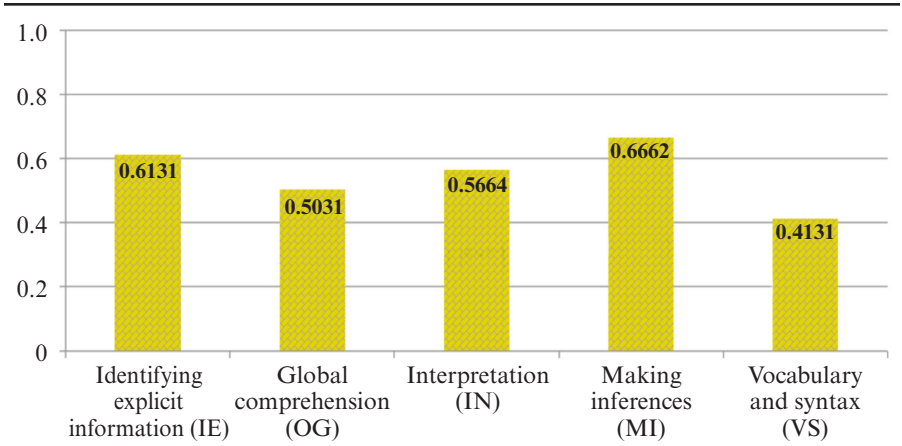

Figure 2. Probabilities of mastering the attributes for all the students.

Discussion

The purpose of this study was to verify the feasibility of modelling the data of 4,762 Canadian students from booklet 13 of PIRLS 2011. Two Q-matrices were developed by three experts. The data were analyzed with DINA and G-DINA to assess the models fit to the data, and to examine the diagnostic quality of the items and the skill mastery profiles of the students.

The fact that the G-DINA model fits the data better than the DINA corroborates the results of the research carried out in mathematics by Basokcu (2014) and by Ma, Iaconangelo and de la Torre (2016). Indeed, the G-DINA relaxes the assumption of equal probability of correct answers when students do not master all the required attributes. Thus, even if they do not master all the attributes, the probability of answering correctly may vary across participants, given the number and type of attributes that are not mastered (Loye, 2010).

Moreover, G-DINA is often less influenced than other specific models when there are changes in the Q-matrices (Basokcu, 2014), which is the case in our study (between the final Q1- and Q2-matrices). Despite the better fit of generalized DCMs, the specific models, when used correctly nevertheless offer the possibility of obtaining simpler and stable interpretations, and provide more accurate attribute mastery profiles (Ma,

Iaconangelo & de la Torre, 2016). One suggestion is to use the Wald test for each item to determine whether a generalized DCM can be replaced by a specific DCM without losing the quality of model fit to the data (Ma, Iaconangelo & de la Torre, 2016). However, we did not perform this step, because it is not the main focus of our study.

The results of the correct proportion and transformed correlation show that the models fit the data adequately, but not according to the log-odds ratios. Thus, the question about the reliability and sensitivity of these statistics arises when the three indicators do not result in the same conclusions (Chen, de la Torre & Zhang, 2013). The research suggests there are probably problems of inaccuracy in the Q-matrix that we need to detect with more sophisticated techniques, such as the Wald test. It is also important to verify which statistic might be reliable in assessing absolute fit depending on the DCM used, the nature of the responses, and the type and number of attributes identified. This point was emphasized in the work of Jang (2005, 2009), Chen, de la Torre and Zhang (2013) and Ma, Iaconangelo and de la Torre (2016).

The model fit is better with the final Q2-matrix than with the Q1-matrix, which emphasizes the multidimensional nature of the reading comprehension, as most items are related to at least two attributes. For the Q1-matrix, all 35 items are one-attribute items, while in the final Q2-matrix, apart from the 9 one-attribute items, there are 26 items with two or more attributes. This fit problem thus highlights the importance of refining the attributes in the Q-matrix, because the more detailed they are, the finer the diagnostic information obtained (Lee & Sawaki, 2009; Li, 2011). This idea is appropriate in the PIRLS 2011 test with the Q1-matrix, as the processes P1 *Examine and evaluate content, language, and textual elements* or P4 *Interpret and integrate ideas and information* require two separated attributes, as explained in our Q-matrix development process. However, a high number of attributes can cause problems with the capacity of DCM modelling, an important factor to consider other than the relevance of attribute mastery profiles.

Thus, a challenge for designers is to balance the number of attributes identified with the length of the test, i.e., more items should be added when a large number of skills is identified for the test (Li, 2011). Sometimes, when the conditions for constructing the Q-matrix are met, the decision

to select the Q-matrix depends on the results provided by the models. In other words, we have to let the DCMs decide which Q-matrix fits the data best, as we did with two Q-matrices.

The guessing and slipping parameters suggest an average diagnostic quality of items, probably because the test was not designed with a diagnostic purpose. This idea has been confirmed in the research of Li (2011), Jang (2009), Ravand, Barati and Widhiarso (2013) and Huang and Wang (2014).

On the other hand, the fact that the diagnostic quality is better in Part 2 of the test can be explained by the link among the reading purposes, the type of text and the psychometric quality of the items. Passage 2 of the test is informational test in nature, with a clearer and more coherent organizational structure than that in Part 1, which is fictional and loosely structured, with conversations between characters. Furthermore, several studies on the influence of textual items and reading comprehension make this observation (Jang, 2009). For example, Freedle and Kostin (1993) report that at least one-third (33%) of the variance in TOEFL RC item difficulty is explained by variables associated with passage content and structure. Alderson, Percsich, and Szabo (2000) maintain that reading proficiency entails the ability to recognize the ideas presented and to understand the author's meanings in a sequence of ideas. Jang (2009) confirms that texts with different rhetorical organizational structures determine students' different cognitive processes.

The MCQs have a higher guessing parameter than constructed-answer questions. However, according to Huang and Wang (2014), this parameter refers not only to item characteristics, but also to student ability, as guessing is the interaction between the tendency of an item to elicit guesses and the student's guessing ability. In addition, proficient students may have a greater ability to guess the answer correctly than less proficient ones. On the other hand, weaker students are easily influenced by distracting factors (Huang & Wang, 2014). These results could be interesting for attribute identification, as textual variables may elicit different cognitive skills, while the choice of question types could influence the diagnostic quality of the items (Jang, 2009).

The probabilities of skill mastery strongly corroborate reading theories and the degree of skill difficulty, as A1 *Identifying implicit information* and A4 *Making inferences* are considered easier than A2 *Global comprehension*

and A3 *Interpretation*. A5 *Vocabulary and syntax* appears to be the most difficult to master, which was reinforced by the finding that lack of vocabulary is the major obstacle to reading comprehension (García, 1991; Jang, 2009; Li, 2011). A basic rule of thumb is that readers must know 95% of the words in a text to read the text successfully (Grabe, 2000; Li, 2011). However, although the attribute A1 *Identifying implicit information* is rated easier than A4 *Making inferences*, it is less mastered by students, by a 5% difference. In our opinion, the complementarity of the attributes in a question contributes to increasing the probability of mastery of the attribute A4 *Making inferences*. Out of 16 items related to A1 *Identifying implicit information*, half of them have a single attribute. Only one item has one attribute among the 12 items for which A4 *Making inferences* has been identified. The remaining 11 items have two, three, four and five attributes.

The most representative profile includes students who master A1 *Identifying explicit information*, A2 *Global comprehension*, A3 *Interpretation* and A4 *Making inferences*, but not A5 *Vocabulary and syntax*. These results correspond well to our expectations, since the *Vocabulary and syntax* attribute is the most challenging, so it is logical that it is the least mastered among students.

The four unlikely student profiles are explained by the compensatory nature of reading skills:

1. Profile 10000, corresponding to students who have mastered only A1 *Identifying explicit information*, is the least represented because half of the questions in which this skill was identified as required are items with two or more attributes, i.e., the student needs at least one other skill to answer the items correctly;
2. Profile 11000 corresponds to students who have mastered only A1 *Identifying explicit information* and A2 *Overall understanding*. This combination is quite rare, as global comprehension is part of understanding implicit information and, in our Q-matrix, is often linked to A4 *Making inferences* or A3 *Interpretation*. Thus, it is unlikely that the student has mastered global comprehension, but not interpretation nor making inferences;
3. Profile 01001, which refers to students who have mastered A2 *Global comprehension* and A5 *Vocabulary and syntax*, is sparse because the A5 global comprehension attribute is often needed to answer

questions about interpretation and to make inferences, but it is used less for global comprehension;

4. Profile 11010 is the one for which students master A1 *Identifying explicit information*, A2 *Global comprehension* and A4 *Making inferences*, but not A3 *Interpretation* nor A5 *Vocabulary and syntax*. This is unlikely because interpretation is always linked to one of the other four skills, so it is rare to master the other three skills without mastering the skill of interpretation.

If we refer to theoretical reading models, we find that the interactive models on which the PIRLS test framework is based highlight this complementary nature of reading skills.

Limitations

The first limitation of this research arises from the fact that the Q-matrix was developed only by the expert panel. We did not have the means to verify the identified attributes with the students' think-aloud protocols. Identifying attributes based primarily on expert suggestions and the PIRLS 2011 test framework could create the problem of attributes overspecifications, as the cognitive processes could differ from what went on in students' mind during testing. In this case, Q-matrix refinement techniques, such as using the Wald test or the empirical method proposed by de la Torre (2009), are recommended to detect the problems of underspecification or overspecification of attributes. This would also be an important direction for future research to improve the models fit to data and the diagnostic quality of PIRLS items.

The second limitation is that the discussion of the diagnostic quality of the PIRLS test items is based on research in mathematics or in reading conducted with other tests, as there is not yet research conducted with the PIRLS from the same perspective.

Conclusion

Despite the average diagnostic quality of the items, which is justified because the test was not initially designed for diagnostic purposes, the modelling results clearly show the possibility of receiving more detailed information on the cognitive strengths and weaknesses of students through

the PIRLS test. The fact that skills are mastered at around 50% argues for the diagnostic potential of the test. In addition, the results suggest that both models (DINA and G-DINA) fit the data adequately with the two Q-matrices. And that they fit better with the final Q2-matrix than the Q1 highlights the multidimensional and complementary nature of reading skills.

Similar research will need to look at refining the Q-matrix, which would improve the diagnostic quality of the PIRLS test items. Detailed profiles of mastered or not-mastered skills with appropriate intervention strategies could be the subject of developing and evaluating diagnostic reports for teachers. To ensure the accuracy of the resulting profiles, typical student profiles should be validated by teachers prior to developing and assessing large numbers of diagnostic reports.

Our research has shown the feasibility of diagnostic modelling of large-scale international test data, such as the PIRLS 2011 test, using DINA and the G-DINA. This research could therefore be used in large-scale national or provincial tests at the elementary level. This would bridge the gap between results of these tests and CDA, with the ultimate goal of supporting students with reading difficulties.

Received: October 3, 2019
Final version: August 6, 2020
Accepted: August 7, 2020

RÉFÉRENCES

Adams, M. J., & Collins, A. M. (1979). A schema-theoretic view of reading. In R. O. Fredolle (Ed.), *Discourse processing: Multidisciplinary Perspectives* (pp. 1-22). Norwood, NJ: Ablex.

Afflerbach, P., & Cho, B. (2009). Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 69-90). New York, NY: Routledge.

Alderson, J. C. (2005). *Assessing reading*. Stutgart, Denmark: Ernst Klett Sprachen.

Alderson, J. C. (2010). "Cognitive diagnosis and Q-matrices in language assessment": A commentary. *Language Assessment Quarterly*, *7*(2), 96-103. doi: 10.1080/15434300903426748

Alderson, J. C., Percsich, R., & Szabo, G. (2000). Sequencing as an item type. *Language Testing*, *17*(4), 423-447. doi: 10.1177/026553220001700403

Alexander, P. A., & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. *Handbook of reading research*, 3, 285-310. doi: 10.4324/9781410605023.ch19

An, S. (2013). Schema theory in reading. *Theory and Practice in Language Studies*, *3*(1), 130-134. doi: 10.4304/tpls.3.1.130-134

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (vol. 1, pp. 255-291). New York, NY: Longman.

Arican, M., & Sen, S. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi/Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 238-253. doi: 10.13140/RG.2.1.1262.5362

Basokcu, T. O. (2014). Classification accuracy effects of Q-matrix validation and sample size in DINA and G-DINA models. *Journal of Education and Practice*, *5*(6), 220-230. Retrieved from https://www.iiste.org/Journals/index.php/JEP/article/view/11253/11543

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119-157. doi: 10.1177/026553229801500201

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423-466. doi: 10.1111/0023-8333.00016

Carrell, P. L. (1983). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a Foreign Language*, *1*(2), 81-92. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl12carrell.pdf

Carrell, P. L., Devine, J., & Eskey, D. E. (Eds.). (1988). *Interactive approaches to second language reading*. Cambridge, UK: Cambridge University Press.

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

Clay, M. M. (1991). *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19-38. doi: 10.1111/j.1745-3984.2011.00158.x

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163-183. doi: 10.1177/0146621608320523

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199. doi: 10.1007/s11336-011-9214-8

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624. doi: 10.1007/s11336-008-9063-2

Desrosiers, H. & Tétreault, K. (2012). *Les facteurs liés à la réussite aux épreuves obligatoires de français en sixième année du primaire: un tour d'horizon*. Québec, QC: Institut de la statistique du Québec. Repéré à http://www.stat.gouv.qc.ca/statistiques/education/prescolaire-primaire/reussite-epreuve-francais.html.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 979-1027). Amsterdam, Netherlands: Elsevier.

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics*, *68*(3), 263-272. doi: 10.1007/s10649-007-9099-8

Dubin, F., Eskey, D. E., Grabe, W., & Savignon, S. (1986). *Teaching second language reading for academic purposes*. Reading, MA: Addison-Wesley.

Evran, D. (2019). An application of cognitive diagnosis modeling in TIMSS: A comparison of intuitive definitions of Q-Matrices. *International Journal of Modern Education Studies*, *3*(1), 4-17. Retrieved from https://www.ijonmes.net/index.php/ijonmes/article/view/33

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. doi: 10.1037/h0031619

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*(2), 134-169. doi: 10.1177/026553229301000203

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. In J. S. Johnson (Ed.), *Spaan fellow working papers in second or foreign language assessment* (vol. 4, pp. 1-39). Ann Arbor, MI: University of Michigan. Retrieved from https://www.researchgate.net/profile/Shudong_Wang4/publication/251842392_Validation_and_Invariance_of_Factor_Structure_of_the_ECPE_and_MELAB_across_Gender/links/0f31753889dcc4ef54000000/Validation-and-Invariance-of-Factor-Structure-of-the-ECPE-and-MELAB-across-Gender.pdf

García, G. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly*, *26*(4), 371-392. doi: 10.2307/747894

Giasson, J. (1996). *La compréhension en lecture*. Bruxelles, Belgique: De Boeck Supérieur.

Gierl, M. J., Cui, Y., & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns to achievement tests. *Journal of Modern Applied Statistical Methods*, *7*(1), 234-245. doi: 10.22237/jmasm/1209615480

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375-406. doi: 10.2307/3586977

Grabe, W. (2000). Reading research and its implications for reading assessment. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 226-262). Cambridge, UK: Cambridge University Press.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois Urbana-Champaign, Champaign, IL.

Huang, H. Y., & Wang, W. C. (2014). The random-effect DINA model. *Journal of Educational Measurement*, *51*(1), 75-97. doi: 10.1111/jedm.12035

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: Linkage to instruction. *Educational Research and Evaluation*, *16*(3), 287-301. doi: 10.1080/13803611.2010.523294

Irwin, J. W. (1991). *Teaching reading comprehension processes*. Englewood, NJ: Prentice-Hall.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (Unpublished doctoral dissertation). University of Illinois Urbana-Champaign, Champaign, IL.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity argument for fusion model application to *LanguEdge* assessment. *Language Testing*, *26*(1), 31-73. doi: 10.1177/0265532208097336

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. doi: 10.1177/01466210122032064

Kasai, M. (1997). *Application of the rule-space model to the reading comprehension section of the Test of English as a Foreign Language (TOEFL)* (Unpublished doctoral dissertation). University of Illinois Urbana-Champaign, Champaign, IL.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163-182. doi: 10.1037/0033-295X.95.2.163

Labrecque, M., Chuy, M., Brochu, P., & Houme, K. (2012). *PIRLS 2011: le contexte au Canada*. Toronto, ON: CMEC.

Langer, J. A. (1990). The process of understanding: Reading for literary and informative purposes. *Research in the Teaching of English*, *24*(3), 229-260. Retrieved from http://www.jstor.org/stable/40171165.

Lee, Y.-S., Johnson, M., Park, J. Y., Sachdeva, R., Zhang, J., & Waldman, M. (2013, April). *A multidimensional scaling (MDS) approach for investigating students' cognitive weakness and strength on the TIMSS 2007 mathematics assessment*. Paper presented at the 2013 Annual Conference of the American Educational Research Association, San Francisco, CA. Retrieved from https://www.researchgate.net/publication/244988418_A_MDS_Approach_for_Investigating_Student's_Cognitive_Weakness_and_Strength_on_the_TIMSS_2007_Mathematics_Assessment

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*(2), 144-177. doi: 10.1080/15305058.2010.534571

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172-189. doi: 10.1080/15434300902985108

Lee, Y.-W, & Sawaki, Y. (2011). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239-263. doi: 10.1080/15434300903079562

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3-16. doi: 10.1111/j.1745-3992.2007.00090.x

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. In J. S. Johnson (Ed.), *Spaan fellow working papers in second or foreign language assessment* (vol. 9, pp. 17-46). Ann Arbor, MI: University of Michigan. Retrieved from https://www.academia.edu/9788619/A_cognitive_diagnostic_analysis_of_the_MELAB_reading_test

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1-25. doi: 10.1080/10627197.2013.761522

Loye, N. (2010). 2010, odyssée des modèles de classification diagnostique (MCD). *Mesure et évaluation en éducation*, *33*(3), 75-98. doi: 10.7202/1024892ar

Loye, N., & Lambert-Chan, J. (2016). Au cœur du développement d'une épreuve en mathématique dotée d'un potentiel diagnostique. *Mesure et évaluation en éducation*, *39*(3), 29-57. doi: 10.7202/1040136ar

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 1-18. doi: 10.1177/0146621615621717

Ma, X., & Meng, Y. (2014). Towards personalized English learning diagnosis: Cognitive diagnostic modelling for EFL listening. *Asian Journal of Education and e-Learning*, *2*(5), 336-348. Retrieved from https://ajouronline.com/index.php/AJEEL/article/view/1669

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Amsterdam, Netherlands: International Association for the Evaluation of Educational Achievement.

Pagani, L. S., Fitzpatrick, C., Belleau, L., & Janosz, M. (2011). Prédire la réussite scolaire des enfants en quatrième année à partir de leurs habiletés cognitives, comportementales et motrices à la maternelle. *Étude longitudinale du développement des enfants du Québec (ÉLDEQ 1998-2010): de la naissance à 10 ans*. Repéré à http://www.jesuisjeserai.stat.gouv.qc.ca/publications/fascicule_reussite_scol_fr.pdf

Ravand, H., Barati, H., & Widhiarso, W. (2013). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing*, *3*(1), 11-37. Retrieved from https://www.researchgate.net/publication/281558416_Exploring_Diagnostic_Capacity_of_a_High_Stakes_Reading_Comprehension_Test_A_Pedagogical_Demonstration

Rumelhart, D. E. (1978). *Schemata: The building blocks of cognition*. San Diego, CA: Center for Human Information Processing, University of California.

Rupp, A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Samuels, S. J., & Kamil, M. L. (1984). 7 Models of the reading process. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 185-224). Mahwah, NJ: Lawrence Erlbaum.

Scott, H. S. (1998). *Cognitive diagnostic perspectives of a second language reading test* (Unpublished doctoral dissertation). University of Illinois Urbana-Champaign, Champaign, IL.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*(4), 461-488. doi: 10.3102/10769986029004461

Snow, C. (2002). *Reading for understanding: Towards an R&D program in reading comprehension*. Santa Monica, CA: Rand.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, *16*(1), 32-71. doi: 10.2307/747348

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287-305. doi: 10.1037/1082-989X.11.3.287

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item-based model fit evaluation for the TIMSS 2011 assessment. *SAGE Open*, 9(1), 1-11. doi: 10.1177/2158244019832684

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Educational Research Journal*, *26*(2), 237-255. doi: 10.1007/s13394-013-0090-7

Toker, T., & Green, K. (2012, April). *An application of cognitive diagnostic assessment on TIMMS-2007 8th grade mathematics items*. Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, BC. Retrieved from https://files.eric.ed.gov/fulltext/ED543803.pdf

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

von Davier, M. (2008). A general diagnostic model applied to language testing data. Princeton, NJ: ETS.

Wafa, M. N. (2019). Assessing school students' mathematic ability using DINA and DINO models. *International Journal of Mathematics Trends and Technology (IJMTT)*, *65*(12), 153-165. Retrieved from http://www.ijmttjournal.org/Volume-65/Issue-12/IJMTT-V65I12P517.pdf

Wafa, M. N., Hussaini, S. A. M., & Pazhman, J. (2020). Evaluation of students' mathematical ability in Afghanistan's schools using cognitive diagnosis models. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(6), em1849. doi: 10.29333/ejmste/7834

Yamaguchi K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLOS ONE*, *13*(2), e0188691. doi: 10.1371/journal.pone.0188691

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 119-145). Cambridge, UK: Cambridge University Press. Retrieved from https://pdfs.semanticscholar.org/32a5/00670eced66b3023e45a8b0929ad4c1f46e3.pdf