Delafontaine *et al.*
Large-scale Assessments in Education (2022) 10:19
https://doi.org/10.1186/s40536-022-00138-4

Large-scale Assessments in Education

RESEARCH Open Access

Using country-specific Q-matrices for cognitive diagnostic assessments with international large-scale data

Jolien Delafontaine[1†], Changsheng Chen[1,2*†], Jung Yeon Park[1,2,3] and Wim Van den Noortgate[1,2]

†Jolien Delafontaine and Changsheng Chen contributed equally to this work as first authors

*Correspondence: changsheng.chen@kuleuven.be

[1] Faculty of Psychology and Educational Science, KU Leuven, Leuven, Belgium
[2] Imec Research Group ITEC, KU Leuven, Kortrijk, Belgium
[3] College of Education and Human Development, George Mason University, Fairfax, USA

Abstract

In cognitive diagnosis assessment (CDA), the impact of misspecified item-attribute relations (or "Q-matrix") designed by subject-matter experts has been a great challenge to real-world applications. This study examined parameter estimation of the CDA with the expert-designed Q-matrix and two refined Q-matrices for international large-scale data. Specifically, the G-DINA model was used to analyze TIMSS data for Grade 8 for five selected countries separately; and the need of a refined Q-matrix specific to the country was investigated. The results suggested that the two refined Q-matrices fitted the data better than the expert-designed Q-matrix, and the stepwise validation method performed better than the nonparametric classification method, resulting in a substantively different classification of students in attribute mastery patterns and different item parameter estimates. The results confirmed that the use of country-specific Q-matrices based on the G-DINA model led to a better fit compared to a universal expert-designed Q-matrix.

Keywords: G-DINA, Q-matrix refinement, Stepwise validation method, Nonparametric classification method, TIMSS 2011 mathematics, International comparison

Introduction

International comparative assessments such as PISA (Programme for International Student Assessment) or TIMSS (Trends in International Mathematics and Science Study) have the power to influence educational policy and practice to a large extent (Sedat & Arican, 2015). Item response theory (IRT; Baker, 2001) has traditionally been used to analyze such large-scale assessments and to provide information about students' abilities. This method summarizes students' overall ability in a particular subject (e.g., mathematics, reading, or science) by means of a single ability score (Chen, 2017; Nájera et al., 2019). Student achievement is then compared across countries and an international benchmark is set. Unfortunately, the general ability score provides neither teachers nor policymakers with the fine-grained diagnostic information necessary to determine if students have mastered a particular domain. This makes the implementation of a targeted educational strategy based on international large-scale assessments difficult. Cognitive diagnosis assessment (CDA) allows to understand students' assessment outcomes

by fine-grained attributes[1] that are directly related to students' success in a given subject domain so that statistical data analysis may provide richer information regarding what types of attributes students have mastered (Jurich & Bradshaw, 2014).

The Generalized Deterministic Inputs, Noisy "And" Gate model (G-DINA; de la Torre, 2011), one of the popularly used psychometric models for the CDAs, can be used for just this purpose. The G-DINA aims to measure to what extent students master a set of cognitive attributes (e.g., fractions, proportions, and decimals as a fine-grained cognitive mathematic attribute) to improve educational policy and practice. Some studies have already employed the G-DINA to analyze data from international comparative tests, such as PISA (Jia et al., 2021; Wu et al., 2020). However, before the G-DINA can be used on international assessments to perform the CDA, a content analysis of the test must occur (von Davier & Lee, 2019). Domain experts conduct this analysis to identify a set of related attributes or skills that measure a few broad domains and to define each item by the subset of attributes (Nájera et al., 2019). Researchers refer to such an internal structure where the item-attribute relations are specified as a "Q-matrix" (Tatsuoka, 1984). This two-dimensional matrix with items and attributes defining rows and columns, respectively, includes only one (the attribute is required to solve the item) or zero (the attribute is not required to solve the item).

Currently, using the Q-matrix for international tests has two major drawbacks. First, the Q-matrix is designed by the judgements of experts. In other words, content experts specify the cognitive attributes and their relations with the items. The expert-designed Q-matrix is not always perfect and it could be possible to have some misspecifications. When the content analysis of the Q-matrix is conducted by the fallible judgements of subject-matter experts (Chen, 2017; Nájera et al., 2019; Terzi & de la Torre, 2018), misspecifications of the Q-matrix could have serious consequences for the estimation of students' attribute patterns and the interpretation of the data consequently (de la Torre & Chiu, 2016; Köhn & Chiu, 2018; Nájera et al., 2019). Additionally, experts from different countries may have disagreements on the relationship between items and attributes because of their different educational backgrounds, country-specific curriculum, and teaching situations. Those disagreements may produce uncertainty about the Q-matrix, which provides space for further improvements as well. Because of those, researchers have proposed various refinement methods (Chiu, 2013). This study addresses the potential impact of a misspecified Q-matrix on the CDA in TIMSS and explores the performance of refined Q-matrices.

Second, the same Q-matrix is specified for every participating country. The question that arises in the literature is that, despite the prior work on this issue, little evidence supports the use of a common Q-matrix for different types of population groups. That is, we cannot simply assume that the expert-designed universal Q-matrix from the TIMSS will perform similarly when used to analyze data from different countries. Further, we seek to determine if a particular refinement approach performs better across various countries.

[1] An attribute can be defined as a *"skill or content knowledge that is required to solve a test item."* (Choi et al., 2015).

In the following section, we describe the conceptual background of the G-DINA and two different Q-matrix refinement methods. Next, we apply these techniques to study whether different Q-matrix refinement approaches come to different solutions within and between countries, and what solution provides the best model fit. Finally, we investigate the usefulness of country-specific Q-matrices.

G-DINA

The G-DINA belongs to the family of Cognitive Diagnostic Models (CDMs; Rupp et al., 2010), which is considered as a special case of Latent Class Models (LCMs; Hagenaars & McCutcheon, 2002) where the attribute patterns are modelled to categorize students by means of latent class variables. To be specific, the attribute patterns of the students are conceptually unobservable and therefore have to be measured by their observed responses to a set of items in a test (Chen, 2017). CDMs are confirmatory models in nature because the relationships between the categorical latent variables (attributes) and the test items are defined a priori in a Q-matrix (Ravand & Robitzsch, 2015). There are a number of different modelling approaches within CDMs that have been in use, depending on how relationships between attributes and item responses are modelled and how attributes themselves are combined (Ravand & Robitzsch, 2015; Rupp et al., 2010).

Many CDMs assume certain relationships between items and attributes, such as the Deterministic, Inputs, Noisy, "And" Gate model (DINA; de la Torre, 2009; Junker & Sijtsma, 2001) which assumes that attributes are conjunctive, the Deterministic, Inputs, Noisy, "Or" Gate model (DINO; Templin & Henson, 2006) which assumes that attributes are disjunctive, and so forth. However, sometimes, the attribute relationship is unclear before the model application. In that respect, the G-DINA seems to be a reasonable choice to fit real-life data because it does not have constraints for the relationship of attributes (i.e., conjunctive, disjunctive, and additive assumption for attributes; de la Torre, 2011).

A saturated G-DINA for dichotomous responses with identity link is expressed as follows (de la Torre, 2011; von Davier & Lee, 2019). The model estimates the probability of success for item j under different attribute patterns. k refers to a required attribute based on Q-matrix and K_j^* is the total number of required attributes for item j. $\boldsymbol{\alpha}_{\boldsymbol{lj}}^*$ denotes the reduced attribute vector for item j,$l = 1, ..., 2^{K_j^*}$, which keeps only required attributes. $P\left(X_j = 1|\boldsymbol{\alpha}_{\boldsymbol{lj}}^*\right)$ denotes the probability of the correct answer to item j conditional on attribute pattern$\boldsymbol{\alpha}_{\boldsymbol{lj}}^*$. For instance, in a test designed for measuring four attributes, answering item j correctly needs the 2nd, 3rd, and 4th attribute (i.e.,$K_j^* = 3$), and the reduced attribute pattern is denoted by$\boldsymbol{\alpha}_{\boldsymbol{lj}}^* = (\alpha_{l2}, \alpha_{l3}, \alpha_{l4})'$.

$$P(X_j = 1|\boldsymbol{\alpha}_{\boldsymbol{lj}}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{1}$$

In addition, δ_{j0} is the intercept indicating the baseline probability (i.e., the success probability without any required attribute being mastered); δ_{jk} is the main effect, namely the change in probability when students master a single attribute α_k; $\delta_{jkk'}$ is the first-order interaction effect due to α_k and $\alpha_{k'}$, indicating the change in probability with the mastery of two attributes; $\delta_{j12\ldots K_j^*}$ is the effect of the mastery of all attributes. The higher-

order interaction effects can be added to the model, but are skipped here due to parsimony (denoted by three dots in the formula). Parameters of the G-DINA are estimated by an expectation–maximization implementation of marginalized maximum likelihood estimation (de la Torre, 2011).

Q-matrix refinement methods

To handle uncertainty about some of the item-attribute relations in the Q-matrix, several studies proposed parametric (de la Torre & Chiu, 2016) and nonparametric (Chiu, 2013; Desmarais & Naceur, 2013) approaches to validate and refine the initial design. It is important to notice that these methods are all confirmatory in nature in the sense that they refine the Q-matrix initially designed by experts (Nájera et al., 2019). Furthermore, these refinement methods only handle misspecifications of how items are linked to the attributes (misspecification of rows in the Q-matrix or q-vectors) and not misspecification of the set of underlying attributes (misspecification of the columns of the Q-matrix) (Chiu, 2013). Lastly, these refinement methods are data-driven, so they use students' answers to the test to construct a refined Q-matrix. Thus, a Q-matrix refined by the same method but using a different dataset (e.g., the TIMSS 2011 8th grade mathematics scores from a different country), is therefore likely to be different.

Stepwise validation method

The stepwise validation method for Q-matrix refinement was proposed by Ma and de la Torre (2020a, b). It combines the G-DINA Discrimination Index (GDI; de la Torre & Chiu, 2016) and the Wald statistic based on the G-DINA given an expert-defined Q-matrix, which can be regarded as an extended version of the GDI method. This method is originally designed for graded response data in tandem with the sequential G-DINA. When dichotomous items are applied, this method can still work based on the G-DINA. Compared to other Q-matrix refinement methods, the stepwise validation method does not need assumptions about the processing function and it can consider the item parameter estimation errors because of including the Wald statistic (Ma & de la Torre, 2020a, b). The mechanism of this method can be simply explained by two steps. The first step is to select required attributes by using GDI and the second step is to compare candidate attributes based on the Wald statistic. The criteria corresponding to GDI for selecting the best-recommended q-vector is Proportion of Variance Accounted For (PVAF). de la Torre and Chiu (2016) recommended 0.95 as the rule of thumb for PVAF. When the PVAF of q-vectors is larger than 0.95, a q-vector with the lowest number of attributes will be recommended.

The operation is explained in detail and a flowchart can be consulted in Appendix 1 (Fig. 3). For a certain item, two sets of attributes are defined, including a set A of all required attributes and a set B of all target (or candidate) attributes that need to be tested. A q-vector search bank C with all single-attribute competing q-vectors is defined as well. Initially, A is an empty set and B is a set with all attributes. The first step is to replace the provisional q-vectors in an expert-designed Q-matrix with the competing q-vector in C for the target attribute in B, and calculate relevant PVAF (Ma & de la Torre, 2020a, b). The target attribute with the highest PVAF is defined as a required attribute, and it will be moved from B to A. The second step is to examine whether the q-vectors

with required attributes from A are recommended by GDI (i.e., PVAF > 0.95; Ma & de la Torre, 2020a, b). When the PVAF of the q-vector with the required attribute from A is higher than 0.95, the validation process will stop and it means the required attribute in A is validated. If not, the search bank C will be updated where the competing q-vector becomes a vector with all required attributes in A and one target attribute in B. There will be some competing q-vectors with all required attributes in A and different target attributes in B simultaneously. Then, the target attributes need to be examined whether they are necessary for competing q-vectors by performing tests based on the Wald statistic. If the Wald statistic suggests all target attributes are not necessary, the validation procedure will stop. Otherwise, at least one target attribute can be recommended. Among them, the one in the competing q-vector with the highest PVAF is regarded as the required attribute, and it will be moved from B to A. After that, the required attribute is examined by the Wald statistic, and the unnecessary one will be moved from A to B. The procedure is iterated until no attributes can be added or removed between set A and set B (Ma & de la Torre, 2020a, b).

Some limitations of this method have been identified. In the proposed algorithm, the cut-off value for PVAF in the part of GDI is fixed to 0.95, which has been criticized. Liu (2015) and Wang et al. (2018) supported that the cut-off value should be adjusted based on sample size. Another limitation is the Wald statistic. As Ma and de la Torre (2020a, b) said, the Wald test is an important component of the stepwise validation method and its performance can be further improved when using a better-estimated variance–covariance matrix (Liu et al., 2019).

Chiu's nonparametric classification method

Chiu (2013) also proposed a method to identify and correct mispecified q-vectors in a Q-matrix. The method is based on the nonparametric classification method and comparisons of the residual sum of squares (RSS) between the observed and predicted responses among all the possible Q-matrices given an expert-based Q-matrix. The algorithm consists of various steps. A flowchart is added in Appendix 1 (Fig. 4) to clarify the different steps of the algorithm (Chiu, 2013). The algorithm begins by selecting the item with the highest RSS, which is most likely to be misspecified, and the q-vector that should be updated. Then, the algorithm searches over all possible q-vectors and replaces the q-vector under consideration with the one with the lowest RSS. The algorithm is an iterative procedure where it will stop when all items are visited and the RSS of each item hardly changes anymore (Chiu, 2013). An advantage of this method, compared to model-based methods, is that it does not rely on the model parameters of CDMs when optimizing the algorithm. Furthermore, the approach guarantees good student classification even when the true CDM underlying the observed item responses is unknown. Performance, effectiveness, efficiency, and applicability were proven through simulation studies by Chiu (2013).

One major limitation of this method is that it is unable to handle missing data in the dataset. Because of the booklet design of many large-scale assessments (including TIMSS), the missingness by design is a common feature of these studies. Thus, we need to impute all the missing data before Chiu's method can be performed. Second, Chiu's method is a nonparametric method while parametric models should provide more

Table 1 Number of (Selected) Students of TIMSS 2011 8th Grade Mathematics Test per Country

Country	Total Students	Selected Students
Finland	4,266	2,397
United States	10,477	5,909
Singapore	5,927	3,069
Australia	7,556	4,240
Tunisia	5,128	2,883
Total	33,354	18,498

Selected students are students that filled in a TIMSS 2011 8th grade mathematics assessment booklet where at least one of the 89 items in the Q-matrix was assessed

powerful results when the distributional assumptions are not violated, especially for large samples (Terzi & de la Torre, 2018).

The overarching goal of this study is to explore the adequateness of (1) the selected Q-matrix refinement techniques and (2) the country specificity of Q-matrices in the case of the TIMSS 2011 8th grade mathematics test. This leads to the following research questions:

- Do the country-specific refined Q-matrices offer a better model fit than the original Q-matrix designed by domain experts? If so, is there a particular refinement method that performs better (or worse) than other methods?
- Does using the country-specific Q-matrices with the best model fit alter the interpretation of diagnostic assessments for the TIMSS 2011 eighth-grade mathematics assessment? If so, does this impact differ across countries?

Materials and methods

TIMSS 2011 8th grade mathematics

For this study, we used data from the TIMSS 2011 eighth-grade mathematics assessment. Specifically, we used student responses to 89 items that were released in the TIMSS database. They were comprised of 48 multiple choice items, 32 open-ended questions, and 9 constructed response questions. A score of 1 was given to a completely correct answer and a score of 0 was given to a partly correct or wrong answer. Omitted items were scored as incorrect (0) and missing items by design were scored as non-available (NA). Two criteria were used to select countries in the database: (1) the TIMSS 2011 results of the country were reliably measured (according to Mullis et al., 2012); and (2) the five countries are part of different continents. Consequently, we chose Finland (Europe), the USA (North America), Singapore (Asia), Australia (Oceania), and Tunisia (Africa) for the study. Table 1 provides the sample sizes from five countries.

Q-matrix for 8th grade TIMSS 2011 mathematics

In order to analyze data with the G-DINA, subject-matter experts must prepare a Q-matrix for the test items. Details about constructing Q-matrices with regard to TIMSS 2007 and 2011 are presented in previous studies (Johnson et al., 2013). In essence, the procedure begins with four content domains specified in the original TIMSS framework (i.e., number, algebra, geometry, and data & chance) where the domains are further

Table 2 Attributes and Frequency of attributes in the Original Q-matrix

Attribute	Frequency
Whole numbers and integers (1)	25
Fractions, decimals and proportions (2)	19
Patterns (3)	12
Expressions, equations and functions (4)	21
Lines, angles and shapes (5)	10
Measurement (6)	8
Location and movement (7)	9
Data organisation, representation and interpretation (8)	21
Probability (9)	5

Attributes and frequency adopted from Johnson et al. (2013)

described by multiple topic areas and the accompanying 55 objectives that are a part of math curricula from a majority of countries. For TIMSS 2011 8th grade mathematics assessment, the experts combined more related objectives and defined a total of nine attributes comprised of 89 items. The list of attributes and the number of items involved in the Q-matrix can be found in Table 2. Descriptions of those attributes are available in Johnson et al. (2013) (see the excerpt in Appendix 2). The Q-matrix for those released item sets in the TIMSS 2011 8th grade mathematics assessment is available in Park et al. (2017).[2] Figure 1 gives an example of a multiple-choice question among the released items. According to the Q-matrix, this item requires mastery of two attributes, *expressions, equations and functions* and *measurement* to be answered correctly.

Analyses

Q-matrix refinements and G-DINA

The research questions were investigated empirically by fitting the G-DINA using the original and refined Q-matrices. First, the expert-designed Q-matrix (for the 89 selected items) was validated and refined based on the five selected countries' data as well as on the combined data from the five countries by two selected refinement methods: (1) stepwise validation method (further referred to as stepwise method) (2) Chiu's nonparametric classification method (further referred to as Chiu's method). Next, the Q-matrices of each of the five countries and of the combined data were compared with the expert-designed Q-matrix. Thereafter, the analysis of the G-DINA with the different Q-matrices (i.e., one expert-designed Q-matrix and two refined Q-matrices based on two refinement methods) was conducted for each country. To avoid the problem of using the same data for refining the Q-matrix and estimating model-fit indices to make relevant evaluations stable and reliable, the data of each country were divided into two parts: a random subset of 50% of the data was used for Q-matrix refinement and other 50% for the G-DINA estimation. This operation was repeated ten times and the average value for each model-fit index was used as the evidence for conclusions. We used R 4.1.2 (R Core Team, 2021) with the G-DINA package (version 2.8.8; Ma & de la Torre, 2020a, b) and the NPCD

[2] The expert-designed Q-matrix can be found in Appendix 3.

package (version 1.0–11; Zheng et al., 2019) to perform the Q-matrix refinements and G-DINA analysis.

Handling missing data

Since Chiu's method is unable to handle missing data, we needed to replace the missing data with substitute values. To this end, the common imputation method of Predictive Mean Matching (PMM; Little, 1988; Rubin, 1986) was used for each column of the data. The basic idea of PMM is that for each missing value, the method forms a small set of candidate values and matches one of those observed values for the corresponding missing cell. By using PMM, all missing entries in the different datasets could be replaced. It is worth noting that most of the missings in the datasets were in fact produced by test design (i.e., booklet format) and the planned missings are considered as missing completely at random, so the analysis based on imputed data does not cause biased results (Little & Rubin, 2002). In order to make relevant estimated results reliable and make the methods comparison fair, the same imputed datasets were administrated for all following Q-matrix comparisons (i.e., the expert-designed Q-matrix and the Q-matrix refined by the stepwise method and Chiu's method).

Model evaluation criteria

A series of analyses were conducted within and across countries to provide answers to two research questions. Note that due to overlap between the two questions, some of the results provide answers to both. To deal with the first research question, model fit was investigated within and across countries. Specifically, Akaike's Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) were used as relative fit indices; the limited-information version of Root Mean Square Error of Approximation ($RMSEA_2$; Maydeu-Olivares & Joe, 2014; Ma & de la Torre, 2020a, b) and Standardized Root Mean Square Residual (SRMSR; Liu et al., 2016; Maydeu-Olivares, 2013) were used as absolute fit indices. As for the rule of thumb for these model-fit indices, lower AIC and BIC values indicate a better fit. Because clear guidelines for evaluating the G-DINA in terms of the $RMSEA_2$ and SRMSR are still lacking in the current literature (von Davier & Lee, 2019), we chose to use "Below 0.05" for SRMSR and "Below 0.045" for $RMSEA_2$ to indicate good model fit. These cut-offs are currently used in relevant research based on IRT models and loglinear cognitive diagnosis models (LCDM) respectively (Liu et al., 2016, 2017; Maydeu-Olivares et al., 2011). First, the best-fitting Q-matrix per country was determined by comparing the relative and absolute model fit criteria between the three different kinds of Q-matrices (i.e., original, stepwise, and Chiu's Q-matrix) within each country. Then, mean rankings of model fit indices were calculated as a result of incorporating an expert-designed Q-matrix and the two refined Q-matrices in the G-DINA across five selected countries. The overall rank was calculated by (1) sorting the three Q-matrix types per country (e.g., 1 = best to 3 = worst for Finland), and then (2) averaging the given numbers over the five countries. That is, the Q-matrix ranked lowest indicates that it was able to provide the best model fit for the G-DINA analysis in general.

Given that different refinement methods were evaluated and the best-fitting solutions were chosen for each country, the second research question focused heavily on looking

at their impact on the interpretation of TIMSS. Student attribute mastery percentages were calculated and compared within and across countries to verify if using different Q-matrices alters the estimation of diagnostic information from the TIMSS assessment in general and for each country separately. The calculation of percentages was based on an attribute mastery status of each student estimated by the G-DINA. Differences in student attribute mastery between the original and the refined Q-matrices larger than 10% were retained as a remarkable difference.

Next, we calculated degrees of (dis-)agreement between the original and the two refined Q-matrices for all attributes. When the estimated student mastery matrices based on two comparable Q-matrices classified a student in the same category (master (1) or non-master (0)) we considered this as an agreement. When they classified a student in a different category (classifying a non-master as a master or classifying a master as a non-master), we considered this as a disagreement. The percentages of mastery/non-mastery agreement were calculated for each pair of Q-matrices for each country separately. As an example, the original Q-matrix was compared to the Q-matrix refined by the stepwise method in terms of the mastery agreement rate for one attribute in a dataset with 100 students. The estimated student mastery matrix from the original Q-matrix can be compared to the one from the refined Q-matrix. To that end, the number of the same mastery entries can be counted (for instance giving a value of 80 if both approaches agreed that 80 students mastered this attribute). The number of the different mastery entries could be observed as well (for instance leading to a value of 20 if it was found that based on the original matrix 20 students mastered this attribute but the result based on the refined Q-matrix suggested they did not master this attribute). Then, the mastery agreement of this attribute between two Q-matrices in this dataset would be 80% (i.e., $80/(80+20)$). This kind of calculation would be applied to each country and each attribute in the following analysis. The relevant results from five countries were combined to give an overall explanation of each attribute. The average results based on nine attributes were applied to present a general overview of differences in student classification between different Q-matrices across countries. Third, differences between Q-matrices concerning the interpretation of item parameter estimates, e.g., intercept parameter that measures baseline effect (i.e., the success probability without any required attribute), were scrutinized.

Results[3]

First, the expert-designed Q-matrix was compared to different refined Q-matrices. The results suggest that the Q-matrices refined based on the combined five-countries data are dissimilar to the expert-designed Q-matrix. The percentage of different entries for the stepwise method is 26.59% and the one for Chiu's method is 15.61%. Also, the Q-matrices refined based on each country's data differ from the expert-designed Q-matrix, with the differential proportion ranging from 2.5% to 22.97%. Additionally, the refined Q-matrix based on the combined five-countries data is dissimilar from the Q-matrix based on each country's data for both the stepwise method and Chiu's method,

[3] In practice, Chiu's method needs to set the condensation rule (i.e., "AND" or "OR") due to the requirement of relevant R packages. Two condensation rules were both tried in the data analysis. Analyses based on "OR" rule did not converge, so the following parts only present results based on "AND" rule.

Table 3 Results of G-DINA Model Fit

Country	Q-matrix	AIC	BIC	$RMSEA_2$	SRMSR
Finland	Original	120,390.5	124,375.4	0.0761	0.0862
	Stepwise	**117,980.4**	**122,176.9**	**0.0706**	**0.0745**
	Chiu	120,915.6	135,526.8	0.0711	0.0773
United States	Original	292,238.3	296,929.5	0.0846	0.1099
	Stepwise	**276,331.3**	**282,622.1**	**0.0691**	**0.0735**
	Chiu	277,763.4	285,927.0	0.0705	0.0811
Singapore	Original	137,450.0	141,628.3	0.0881	0.1205
	Stepwise	**125,654.6**	**131,715.5**	0.0663	**0.0795**
	Chiu	126,544.2	133,276.5	**0.0651**	0.0889
Australia	Original	210,511.1	214,942.2	0.0851	0.1203
	Stepwise	198,039.7	**203,403.4**	**0.0689**	**0.0766**
	Chiu	**197,796.4**	206,006.8	0.0706	0.0784
Tunisia	Original	138,141.2	142,270.6	0.0814	0.0935
	Stepwise	133,381.0	**138,004.0**	0.0733	0.0796
	Chiu	**133,287.1**	140,533.2	**0.0712**	**0.0793**
Five-countries data	Original	948,582.3	954,166.9	0.0643	0.0973
	Stepwise	**890,001.8**	**901,176.7**	**0.0380**	**0.0427**
	Chiu	914,767.5	927,584.2	0.0461	0.0668

AIC Akaike Information Criterion, *BIC* Bayesian Information Criterion, $RMSEA_2$ the limited-information version of Root Mean Square Error of Approximation, *SRMSR* Standardized Root Mean Square Residual. The best fitting model is in boldface

and the percentage of different entries ranges from 12.11% to 28.21%. Overall, it can be confirmed that the expert-designed Q-matrix is different from the refined Q-matrices, and those differences are worth exploring further.

Goodness of fit

Goodness of fit within countries

Considering that Chiu's method was applied under the conjunctive assumption, in order to make the methods comparison fairer, the interaction parameters of the G-DINA and the model comparison between the G-DINA and the DINA were scrutinized. We find that most of the interaction parameters in the G-DINA are not zero and not close to zero when the expert-designed Q-matrix and the refined Q-matrix from the stepwise method or Chiu's method are applied. Furthermore, the DINA and G-DINA were compared by model-fit indices and the likelihood ratio test based on each country's data was analyzed. The results of model-fit indices present that almost all estimates support the G-DINA across five countries. Only the estimated SRMSR for the Finland data recommends the DINA. All likelihood ratio tests support the G-DINA. Hence, it can be confirmed that using the G-DINA for the model comparison is acceptable.[4]

Table 3 gives the average AIC, BIC, $RMSEA_2$, and SRMSR of the G-DINA for each refinement method per country. According to the relative fit indices, the Q-matrix refined by the stepwise method appears to be the best fitting Q-matrix regardless of the criterion used for four out of the five selected countries (except Tunisia). For Tunisia we see a difference, the BIC value identifies the Q-matrix refined by the stepwise method

[4] Relevant results can be consulted by the following link "https://github.com/supplement-material/Q-matrix-paper".

Table 4 Mean ranking of relative and absolute model fit indices per q-matrix across countries

Q-matrix	Mean Ranking	
	Relative Fit Indices (AIC and BIC)	Absolute Fit Index ($RMSEA_2$ and SRMSR)
Original	2.83	3.00
Stepwise	1.17	1.25
Chiu	2.00	1.75

AIC Akaike Information Criterion, *BIC* Bayesian Information Criterion, $RMSEA_2$ the limited-information version of Root Mean Square Error of Approximation, *SRMSR* Standardized Root Mean Square Residual

as the most appropriate one, but the estimated values of other indices support Chiu's method. None of the model-fit indices supports the original expert-designed Q-matrix. Interestingly, none of the G-DINA yielded a value lower than the cut-off value for the good model fit of 0.045 ($RMSEA_2$) and 0.05 (SRMSR). Most of them are between 0.06 and 0.12. Nevertheless, it is important to keep in mind that these cut-off values apply for LCDM and IRT models, and are not yet evaluated for the G-DINA. To make the conclusion of methods comparison solid, the five-countries-combined data was applied as well. The results indicates that fitting the G-DINA based on the Q-matrix refined by the stepwise method can produce a better model-fit evaluation than others, and the refined Q-matrix is always better than the expert-designed Q-matrix. Those findings are consistent with the results based on each country's data.

Goodness of fit across countries

The mean ranking of relative and absolute model fit indices for the different Q-matrices can be found in Table 4. The relative fit indices identify the Q-matrix refined by the stepwise method as the most suitable one across countries (mean rank = 1.17), followed by the Q-matrix refined by Chiu's method. Across all selected countries, the original Q-matrix is the least preferred. The absolute fit indices suggest the Q-matrix refined by the stepwise method is the best one (mean rank = 1.25) as well, followed by the Q-matrix refined by Chiu's method and the universal expert-designed Q-matrix, which is the same as the results of mean rank based on the relative fit indices.

Student attribute mastery

Table 5 provides an overview of student mastery per attribute, per Q-matrix, and per country. Percentages in this table can be interpreted as follows: according to the original Q-matrix, 38.46% of the Finnish students have mastered the "*Whole numbers and integers*" attribute. To investigate differences between Q-matrix refinement approaches, we compared the percentages of the best fitting Q-matrix (i.e., Q-matrix refined by the stepwise method) according to the model-fit indices (indicated in bold in Table 5, see Goodness of fit) with the percentages of the original Q-matrix. We considered a difference in student mastery percentages of 10% or more (compared to the original percentages) as a remarkable difference. These percentages were indicated by an *.

First, we find the most notable differences between the best-fitting and the original Q-matrix for Singapore. Six attributes show differences in attribute mastery larger than

Table 5 Percentages of attribute mastery using expert-designed and refined Q-matrices

Attribute	Q-matrix	Student Mastery (%)				
		Finland	USA	Singapore	Australia	Tunisia
Whole numbers and integers	Original	38.46	53.34	56.57	46.46	31.49
	Stepwise	32.54*	51.40	75.50*	46.04	35.48*
	Chiu	43.55*	49.30	65.36*	52.50*	28.13*
Fractions, decimals and proportions	Original	45.64	55.90	68.78	49.01	49.53
	Stepwise	27.45*	36.83*	68.43	49.95	50.23
	Chiu	67.88*	70.69*	70.58	45.35	64.00*
Patterns	Original	31.46	26.23	41.22	45.28	45.72
	Stepwise	32.17	26.99	47.02*	30.66*	13.84*
	Chiu	36.55*	22.32*	55.03*	84.91*	35.17*
Expressions, equations and functions	Original	52.86	52.50	61.97	37.29	42.00
	Stepwise	56.90	51.40	47.93*	33.28*	39.40
	Chiu	44.39*	59.10*	58.98	40.21	42.46
Lines, angles and shapes	Original	40.88	23.34	40.34	39.65	30.45
	Stepwise	30.50*	14.35*	44.51*	39.60	34.65*
	Chiu	33.08*	34.29*	46.82*	33.30*	36.91*
Measurement	Original	43.93	42.44	59.20	42.52	25.95
	Stepwise	42.34	33.69*	50.64*	37.00*	26.88
	Chiu	39.67	29.23*	63.08	26.04*	32.54*
Location and movement	Original	53.98	41.72	53.99	43.80	47.90
	Stepwise	57.53	41.78	47.61*	45.52	43.77
	Chiu	72.01*	59.59*	60.05*	53.77*	34.24*
Data organisation, representation and interpretation	Original	75.84	51.41	53.63	35.50	18.35
	Stepwise	75.59	57.52*	52.00	36.37	23.10*
	Chiu	59.87*	44.71*	69.99*	26.11*	35.48*
Probability	Original	58.53	35.67	63.34	58.47	48.28
	Stepwise	58.41	36.59	63.54	49.01*	45.61
	Chiu	56.74	46.96*	47.67*	60.90	33.71*

* difference in student attribute mastery between the original and the refined Q-matrix greater than 10%

10% when comparing the original with the stepwise method's Q-matrix. Except for "*Fractions, decimals and proportions*", "*Data organisation, representation and interpretation*", and "*Probability*", other attributes all show significant differences in percentages with a range from − 22.66% ("*Expressions, equations and functions*"; (47.93–61.97)/61.97 in Table 5) to + 33.46% ("*Whole numbers and integers*"). Second, four attributes show large differences in student attribute mastery between the stepwise and original Q-matrix in the case of the USA, Australia, and Tunisia. For the USA this encompasses the attributes: "*Fractions, decimals and proportions*" (− 34.11%), "*Measurement*" (− 20.62%), "*Lines, angles and shapes*" (− 38.52%), and "*Data organization, representation and interpretation*" (+ 11.89%). For Australia, it includes the attributes: "*Expressions, equations and functions*" (− 10.75%), "*Patterns*" (− 32.29%), "*Measurement*" (− 12.98%), and "*Probability*" (− 16.18%). The results of Tunisia data show remarkable differences for "*Whole numbers and integers*" (12.67%), "*Patterns*" (− 69.73%), "*Lines, angles and shapes*" (13.79%), and "*Data organization, representation and interpretation*" (+ 25.89%). Third, for Finland, large differences in student attribute mastery between the best-fitting (the stepwise Q-matrix) and the original Q-matrix are found for three out of

nine attributes with a range from − 39.86% ("*Fractions, decimals and proportions*") to − 15.39% ("*Whole numbers and integers*").

From Table 5 we can also derive remarkable student attribute mastery differences between the original and the Q-matrix of Chiu's method across countries. Overall, we see a strong agreement between the original, stepwise, and Chiu Q-matrices. Yet, we find 21 remarkable differences between the original and the stepwise Q-matrix and 35 noticeable differences between the original and Chiu's Q-matrix across all countries and attributes. Although we find many differences between the original and the stepwise Q-matrix, we cannot distinguish tendencies for specific attributes. Whether the model based on the stepwise Q-matrix classifies more or less students as masters for a specific attribute depends on the country.

Overall classification of students as masters or non-masters

Table 6 presents the range of agreement rates in the classification of students as masters or non-masters of nine attributes between three different kinds of Q-matrices (original, stepwise, and Chiu's Q-matrix) for the five selected countries. The average percentage of nine attributes regarding the mastery or non-mastery agreement rate gives a general impression for three Q-matrices. First, we see that the original Q-matrix has high agreement rates with the Q-matrices refined by the stepwise method. The agreement rates of mastery and non-mastery are both over 80%. In contrast, agreement rates between Chiu's method and other methods are smaller (mostly between 70 and 75%). Nevertheless, the three different Q-matrices agree on the classification of most of the students as masters or non-masters with rates higher than 70%. In addition, we tried to explore the reason for the inference consequences. The reason could be the overall Q-matrix misfit or the across-countries differences. The Q-matrix refined based on the five-countries data was included in the comparison to clarify this interesting question. The results in Appendix 4 indicate that both could contribute to the inference consequences. Currently, there is no clear pattern.

Estimated item parameters

Figure 2 shows five scatter plots (corresponding to the five countries) that represent the item parameter estimates of the G-DINA, specifically the intercept estimate ($=\widehat{\delta}_{j0}$) where the original Q-matrix (x-axis) and the two refined Q-matrices (y-axis) are used. Recall that the intercept parameter refers to the probability of correctly solving the item j without mastering any required attribute(s) (i.e., the baseline effect). For each plot, the estimates (=dots) that involve the stepwise and Chiu's methods are colored in blue and red, respectively. The dots that fall on the straight line in the plot suggest that the estimates between the original and the refined ones are an exact match. The results show that the intercept estimates using the stepwise method are more similar to estimates of the original Q-matrix than Chiu's method, especially for Finland, Singapore, Australia, and Tunisia. The estimates using Chiu's method (in red) are in general (but more noticeably for Singapore, USA, and Australia) lower than the estimates using the original one and, therefore, appear to deviate remarkably around the bottom right corner. Interestingly, our findings in this figure align with the previous classification results (Table 6) where the agreement rates between the original Q-matrix and the stepwise Q-matrix is larger.

Table 6 Mastery/non-mastery agreement rates (%) for student mastery between each pair of Q-matrices

	Whole numbers and integers		Fractions, decimals and proportions		Patterns		Expressions, equations and functions		Lines, angles and shapes	
	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery
Original vs. Stepwise	92.46	89.03	67.41	79.37	68.83	92.96	83.41	90.89	82.28	95.48
Original vs. Chiu	86.31	84.70	81.87	58.27	75.75	71.41	80.17	78.49	49.89	70.34
Stepwise vs. Chiu	82.50	84.18	87.05	56.60	78.36	67.81	78.96	73.74	51.20	70.07
	Measurement		Location and movement		Data organisation, representation and interpretation		Probability		Average of Nine Attributes	
	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery
Original vs. Stepwise	63.58	82.37	92.32	94.80	95.84	91.63	76.60	80.80	80.30	88.59
Original vs. Chiu	68.62	88.50	78.28	63.60	78.67	83.54	62.84	64.11	73.60	73.66
Stepwise vs. Chiu	57.69	76.94	78.03	62.75	75.02	83.21	69.52	69.13	73.15	71.60

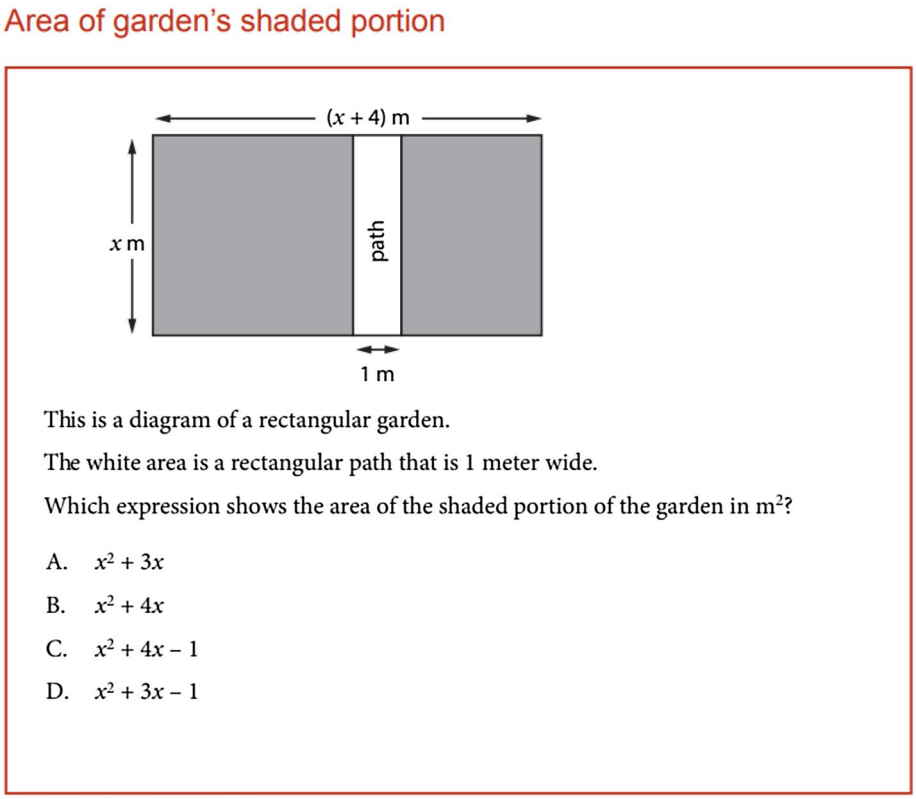
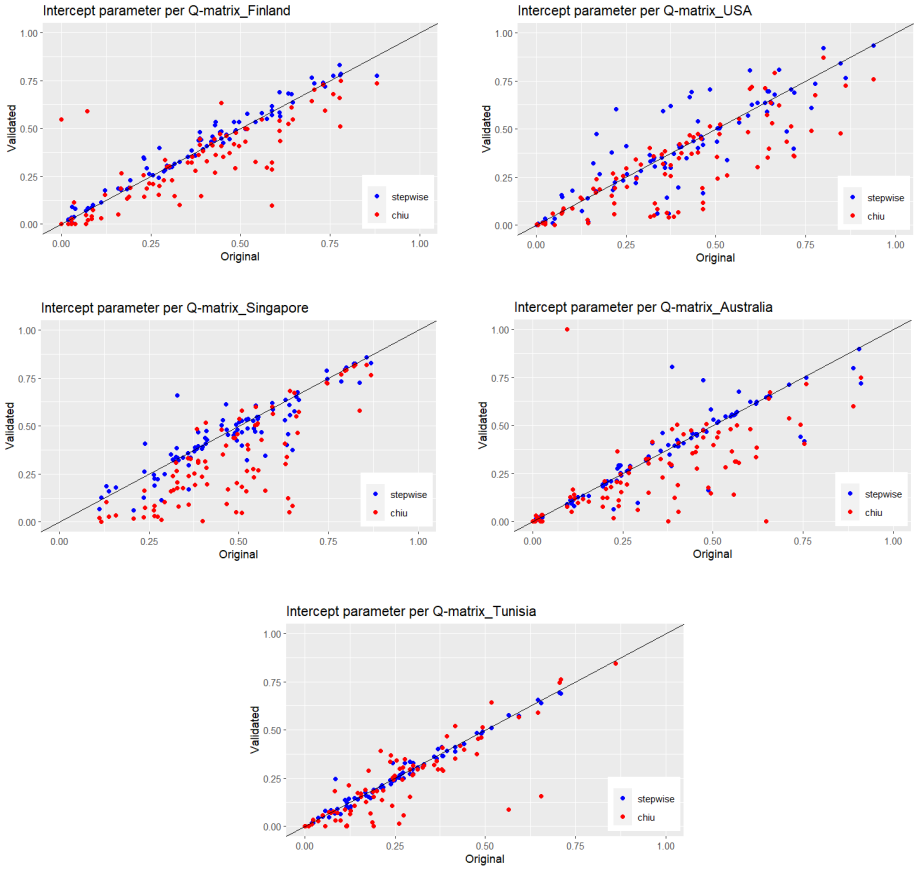


Fig. 1 Example item of the TIMSS 2011 8th grade mathematics test (Mullis et al., 2012)

Discussion

In this study, we examined the impact of a misspecified expert-designed Q-matrix for CDA with the application of international large-scale data (TIMSS 2011 8th grade mathematics). Specifically, our study paid particular attention to the recognition of different Q-matrices for an assessment that was refined by differentiated data structures by countries. First, the performance of the G-DINA using refined Q-matrices as compared to the expert-design Q-matrix was examined with regard to model fit criteria. Our investigation of the TIMSS data made clear that the original Q-matrix that was designed by experts without regard to countries failed to produce an equally good model fit as the two refined Q-matrices: the stepwise validation method (Ma & de la Torre, 2020a, b) or Chiu's nonparametric classification method (Chiu, 2013). This finding, to some extent, naturally justifies the use of the country-specific Q-matrix that was refined by each country separately. While there are equally pros and cons (as mentioned in "Q-matrix refinement methods" section) of the two refinement methods that researchers and practitioners must consider, we found from the TIMSS data that the stepwise method suggested a better model fit than Chiu's method across the countries. The refined Q-matrices by the stepwise method for the five selected countries are provided in Appendix 5. Next, we found that using the G-DINA with the stepwise Q-matrix was noticeably different from the expert-designed Q-matrix in terms of probabilities of



Note. Red = intercept parameters based on Chiu Q-matrix; Blue = intercept parameters based on Stepwise Q-matrix.

Fig. 2 Intercept Parameters per item for the four G-DINA models per Country

attribute mastery, classification accuracy, and item parameter estimation. Furthermore, the impact of using the refined Q-matrices varied across countries.

Several of our findings merit further discussion. First, researchers and practitioners need to consider the advantages and disadvantages of using country-specific Q-matrices. One possible advantage of using it is that the expert-designed (or original) Q-matrix is refined specifically by each country separately. Therefore, any difference in the refined outcomes (i.e., number of attributes required for the item) among countries for an assessment could suggest the country's unique instructional contents. On the other hand, one disadvantage of using the country-specific Q-matrix for CDA for international comparison studies is that it may be unfair (or biased) to directly compare student attribute mastery across countries because the Q-matrix applied in the fitted model was different from country to country after the country-specific refinement, which produced unfair conditions for the across-countries comparison. Another possible disadvantage is that regarding retrofitting CDMs and refining the Q-matrix, the design of large-scale assessments may not satisfy the completeness of the Q-matrix or identifiability conditions, which are required for identifying

proficiency classes and estimating model parameters (Köhn & Chiu, 2016). Overall, we believe that consideration of the population heterogeneity for Q-matrix refinement is a relatively new and unexplored topic in the area; further research is needed for the appropriate use of the refined Q-matrix in real-life assessments.

Notwithstanding the unique insights offered by the current study, there remain some limitations to be considered. First, the two Q-matrix refinement approaches are still contingent on the original Q-matrix designed by experts with respect to the q-entries replacements. Therefore, altering the number of attributes specified in the Q-matrix is beyond the scope of the current study. Furthermore, it is important to notice that when the q-entries in the Q-matrix change, the definition, and interpretation of the attributes may change to some degree. In this way, the best fitting Q-matrix may not necessarily be interpretable or have practical value (Bradshaw et al., 2014; de la Torre, 2008). Moreover, no consensus exists in the current CDA and G-DINA literature regarding the amount and content of specified attributes in the TIMSS mathematics Q-matrix. Researchers specify different Q-matrices for the same test and compare countries according to the predefined attributes (e.g., Im & Park, 2010; Park et al., 2017; Sedat & Arican, 2015). Therefore, a universal attribute design of the Q-matrix remains a critical issue within the current CDA literature (Groß et al., 2016).

Second, the original Q-matrix used in this study was established after the test items in TIMSS were calibrated by a unidimensional IRT. Retrofitting CDMs to the data may result in an unbalanced Q-matrix where some attributes are measured significantly more than others (Sedat & Arican, 2015). In this study, this was most pronounced for the 'probability' attribute that was only measured by five items in the original Q-matrix. This imbalance can distort the attribute classification of students because a small number of items per attribute can generate a situation in which responses to one or a couple of items determines the student's mastery of that attribute (Jurich & Bradshaw, 2014). Therefore, if we want to increase the validity and reliability of student attribute mastery patterns estimated by CDMs, we recommend defining a relevant set of attributes first and then writing items that tap these attributes instead of the other way around (Birenbaum et al., 2005; Bradshaw et al., 2014).

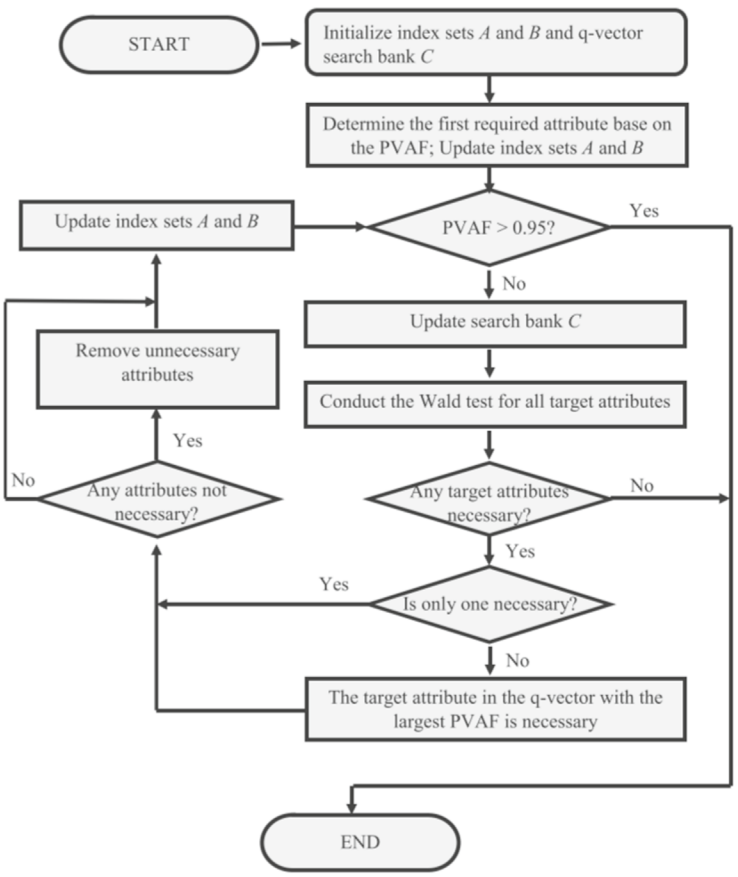
Third, this study is explorative in nature, and some approaches to investigating differences between Q-matrices are very crude. For example, no reliability estimates of student attribute mastery estimates (e.g., Sessoms & Henson, 2018) are provided, nor did we perform any significance tests to investigate differences in student attribute mastery between different Q-matrices. In addition, we do not include relevant domain experts from a specific country to examine whether the refined country-specific Q-matrices recommended by the stepwise method and Chiu's method are meaningful or better than the original universal Q-matrix. We want to stress that the data-driven results are sample-dependent and generalization to other countries or time points may not be warranted. Moreover, they should be used on sufficiently large datasets, but it is not clear yet when a sample size can be considered sufficiently large. Therefore, we recommend the data-driven results to be double-checked in tandem with domain experts, or be used as ancillary information when the Q-matrix is discussed somehow. Relying on the data-driven results solely and blindly is certainly not recommended.

Conclusions

This study provided useful insights concerning differences between Q-matrix refinement techniques, the country-specificity of Q-matrices, and the consequences for practice. Findings from this study could help optimize the Q-matrix so CDMs (e.g., G-DINA) can be more widely used to extract diagnostic attribute-level information out of international comparative tests. Together with the expertise of domain experts, teachers and policymakers could use this fine-grained information to tailor their instruction to students' specific weaknesses and link this diagnostic information to the existing curricula and instructional practices of their particular country. In this way, CDMs are a crucial diagnostic information source that could help improve education systems all over the world.

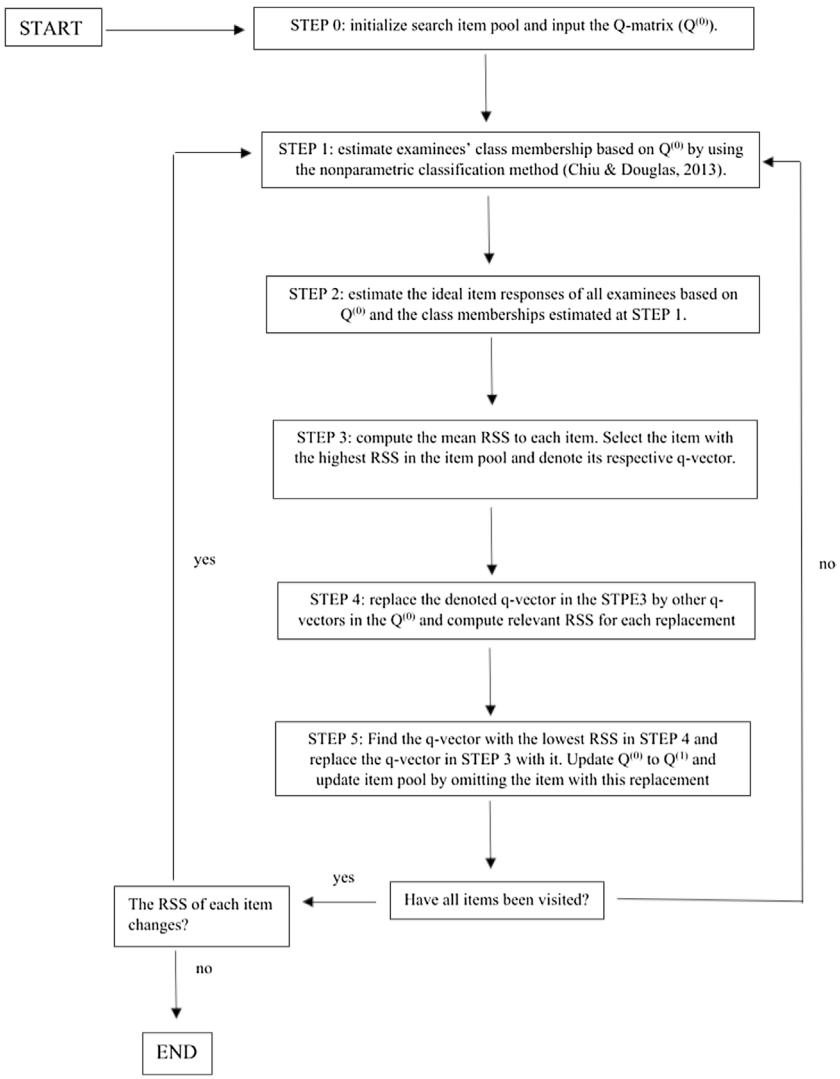
Appendix 1: Flowcharts of the Stepwise Method and Chiu's Method

See Figs. 3, 4.



Note. This flowchart was reprinted from "An empirical Q-matrix validation method for the sequential generalized DINA model." by Ma, W., and De la Torre, J., 2020, British Journal of Mathematical and Statistical Psychology, 73(1), 142–163.

Fig. 3 Flowchart of the Stepwise Method. This flowchart was reprinted from Ma and Torre (2020b)



Note. This flowchart was created based on the explanations of "Statistical refinement of the Q-matrix in cognitive diagnosis." by Chiu, C. Y., 2013, Applied Psychological Measurement, 37(8), 598-618.

Fig. 4 Flowchart of Chiu's method. This flowchart was created based on the explanations of Chiu (2013)

Appendix 2: Attributes of the Original Q-matrix

See Table 7.

Table 7 Attributes, their explanation and frequency in the expert-designed q-matrix

Content Domain	Attribute	Explanation	Frequency
Number	Whole numbers and Integers (1)	Items that asses students' understanding of place value, meaning of operations, multiples/factors, primes, properties, powers and square roots of perfect squares	25
	Fractions, decimals and proportions (2)	Items that asses students' abilities to compare and order decimals and fractions; recognize and compute decimal place value; recognize of compute equivalence; convert between fractions, percents, and decimals; and compute with fractions, decimals, or percents. Also includes items that require students to either reason about or use ratios or proportions. This includes items requiring the formation of ratios; recognizing or determining equivalent ratios; dividing a quantity in a given ratio; and recognizing and determining proportional relationships	19
Algebra	Patterns (3)	Items that require students to extend numeric, algebraic, or geometric patterns or sequences; find missing terms; generalize pattern relationships in a sequence or between adjacent terms or between the sequence number of the term	12
	Expressions, equations and functions (4)	Items that ask students to reason about, solve, or simplify equations of various kinds	21
Geometry	Lines, angles and shapes (5)	Items that require students to understand or apply their understanding of basic geometry in one, two, or three dimensions	10
	Measurement (6)	Items that require students to actually measure or calculate measure such as area of perimeter	8
	Location and movement (7)	Items related to graphing as well as those that require students to visualize and rotate objects, nets or other representations in space	9
Data and chance	Data organisation, representation and interpretation (8)	Items related to creating, interpreting and predicting form data	21
	Probability (9)	Items that reason about simple probability situations like judging the chance of an outcome as certain, more likely, equally likely, less likely, or impossible; use data from experiments or given probabilities to calculate chances or predict future outcomes; and determine chance of particular outcomes	5

Attributes and frequency adopted from Johnson et al. (2013)

Appendix 3: Q-matrix for 89 selected items of TIMSS 2011

See Table 8.

Table 8 Q-matrix designed by experts for 89 Items of TIMSS 2011 8th grade mathematics test

Block	Item	Attribute								
		1	2	3	4	5	6	7	8	9
M01	1	1	0	0	0	0	0	0	0	0
M01	2	0	0	1	0	0	0	0	1	0
M01	3	1	0	1	0	0	0	0	1	0
M01	4a	1	0	1	0	0	0	0	1	0
M01	4b	1	0	1	0	0	0	0	1	0
M01	4c	1	0	1	0	0	0	0	1	0
M01	5	0	0	1	1	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0
M01	8	0	1	0	0	0	0	0	1	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	0	0	0	0	0	0	0	0
M02	4	0	1	0	0	0	0	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	0	0	0	1	0	1	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	1	0	0	1	0	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	1	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	0	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	1	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	1	0	1	0	0	0	0	0	0
M03	10	1	0	0	1	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	1	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0
M03	17	0	1	0	0	0	0	0	1	0
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	0	0	0	0	0	0

Table 8 (continued)

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0
M05	6	1	0	0	1	0	0	0	0	0
M05	7	0	1	0	0	1	0	0	0	0
M05	8	1	0	0	0	0	1	0	0	0
M05	9	0	0	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	0	0	0	0	0	1	0	0
M05	12	0	0	0	0	0	0	1	0	0
M05	13	0	1	0	0	0	0	0	1	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	0	0	0
M06	2	0	1	0	0	0	0	0	0	0
M06	3	1	0	0	0	0	0	0	1	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	0	0	0	0	0	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	0	0	1	0	0	0	0	0	0
M06	6	1	0	0	1	0	0	0	0	0
M06	7	0	0	0	1	0	0	0	0	0
M06	8	0	0	0	1	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	1	0
M06	10B	0	0	0	0	1	0	1	1	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	0	1	0
M07	1	1	0	1	0	0	0	0	1	0
M07	2	0	1	0	0	0	0	0	0	0
M07	3	0	1	0	1	0	0	0	0	0
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	0	0
M07	6	0	0	0	0	1	0	1	0	0
M07	7	1	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	1	0	0
M07	9	1	0	0	1	0	0	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	0	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	0	0	0	0	0	0	1	0

Adopted from Park et al. (2017)

Appendix 4: Mastery/Non-mastery Agreement Rates (%) between each pair of Q-matrices

See Table 9.

Table 9 Mastery/non-mastery agreement rates (%) between each pair of Q-matrices based on five-countries and country-specific data

	Whole numbers and integers		Fractions, decimals and proportions		Patterns		Expressions, equations and functions		Lines, angles and shapes	
	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery
Original vs. Stepwise(cs)	92.46	89.03	67.41	79.37	68.83	92.96	83.41	90.89	82.28	95.48
Original vs. Chiu(cs)	86.31	84.70	81.87	58.27	75.75	71.41	80.17	78.49	49.89	70.34
Stepwise(cs) vs. Chiu(cs)	82.50	84.18	87.05	56.60	78.36	67.81	78.96	73.74	51.20	70.07
Original vs. Stepwise(fc)	87.85	80.11	81.83	60.69	59.54	88.02	75.56	82.65	56.73	65.32
Original vs. Chiu(fc)	87.28	86.37	64.27	49.52	54.34	48.68	82.65	87.49	44.61	72.21
Stepwise(cs) vs. Stepwise(fc)	83.28	85.00	71.21	65.09	59.76	53.79	74.21	80.73	55.75	79.41
Chiu(cs) vs. Chiu(fc)	85.18	80.57	75.12	48.55	81.59	92.62	72.23	76.15	58.94	65.37
	Measurement		Location and movement		Data organisation, representation and interpretation		Probability		Average of Nine Attributes	
	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery	Mastery	Non-mastery
Original vs. Stepwise(cs)	63.58	82.37	92.32	94.80	95.84	91.63	76.60	80.80	80.30	88.59
Original vs. Chiu(cs)	68.62	88.50	78.28	63.60	78.67	83.54	62.84	64.11	73.60	73.66
Stepwise(cs) vs. Chiu(cs)	57.69	76.94	78.03	62.75	75.02	83.21	69.52	69.13	73.15	71.60
Original vs. Stepwise(fc)	74.64	67.72	76.57	79.76	87.71	76.63	68.41	47.66	74.32	72.06
Original vs. Chiu(fc)	74.99	89.90	82.76	85.54	83.51	86.83	60.41	49.65	70.54	72.91
Stepwise(cs) vs. Stepwise(fc)	75.98	83.48	68.80	82.08	77.68	80.77	62.31	51.33	69.89	73.52
Chiu(cs) vs. Chiu(fc)	66.37	59.05	78.21	80.30	86.17	78.41	80.06	57.76	75.99	70.98

"cs" referred to using the country-specific data for that method, and "fc" referred to using the five-countries-combined data for that method

Appendix 5: Refined Q-matrix for five selected countries based on the stepwise method

See Tables 10, 11, 12, 13, 14.

Table 10 Q-matrix refined by the stepwise method: Finland

Block	Item	Attribute								
		1	2	3	4	5	6	7	8	9
M01	1	1	0	0	0	0	0	0	0	0
M01	2	0	0	0	0	0	0	0	1	0
M01	3	1	0	1	0	0	0	0	1	0
M01	4a	0	0	1	0	0	0	0	0	0
M01	4b	1	0	1	0	0	0	0	0	0
M01	4c	0	0	1	0	1	0	0	0	0
M01	5	0	0	1	1	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0
M01	8	0	1	0	0	0	0	0	1	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	1	0	0	0	0	0	0	0
M02	4	0	1	0	0	0	0	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	1	0	0	0	0	0	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	1	0	0	1	0	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	0	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	0	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	0	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	1	0	1	0	0	0	0	0	0
M03	10	1	0	0	1	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	0	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0

Table 10 (continued)

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M03	17	0	1	0	0	0	0	0	1	0
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	0	0	0	0	0	0
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0
M05	6	1	0	0	1	0	0	0	0	0
M05	7	0	1	0	0	1	0	0	0	0
M05	8	1	0	0	0	0	1	0	0	0
M05	9	0	0	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	0	0	0	0	0	1	0	0
M05	12	0	0	0	0	0	0	1	0	0
M05	13	0	0	0	0	0	0	0	1	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	0	0	0
M06	2	0	1	0	0	0	0	0	0	0
M06	3	1	0	0	0	0	0	0	1	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	1	0	0	0	0	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	0	0	1	0	0	0	0	0	0
M06	6	0	0	0	1	0	0	0	0	0
M06	7	0	0	0	1	0	0	0	0	0
M06	8	0	0	0	1	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	1	0
M06	10B	0	0	0	0	0	0	1	1	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	0	1	0
M07	1	1	0	1	0	0	0	0	1	0
M07	2	0	1	0	0	0	0	0	0	0
M07	3	0	1	0	0	0	0	0	0	0
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	0	0
M07	6	0	0	0	0	1	0	1	0	0
M07	7	0	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	1	0	0
M07	9	1	0	0	0	0	0	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	0	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	0	0	0	0	0	0	1	0

Table 11 Q-matrix Refined by the Stepwise Method: United States

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M01	1	1	0	0	0	0	0	0	0	0
M01	2	0	0	1	0	0	0	0	0	0
M01	3	1	0	1	0	0	0	0	1	0
M01	4a	0	0	1	0	0	0	0	1	0
M01	4b	1	0	1	0	0	0	0	1	0
M01	4c	0	0	1	0	0	0	0	0	0
M01	5	0	0	1	1	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0
M01	8	0	1	0	0	0	0	0	0	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	0	0	0	0	0	0	0	0
M02	4	0	1	1	0	0	0	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	0	0	0	1	0	1	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	0	0	1	0	1	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	1	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	0	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	1	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	1	0	1	0	0	0	0	0	0
M03	10	1	0	0	1	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	1	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0
M03	17	0	1	0	0	0	0	0	1	0
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	0	0	0	0	0	0
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0

Table 11 (continued)

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M05	6	1	0	0	1	0	0	0	0	0
M05	7	0	1	0	0	1	0	0	0	0
M05	8	0	0	0	0	0	1	0	0	0
M05	9	0	1	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	1	0	0	0	0	1	0	0
M05	12	0	1	0	0	0	0	1	0	0
M05	13	0	1	0	0	0	0	0	0	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	0	0	0
M06	2	0	1	0	0	0	0	0	1	0
M06	3	1	0	0	0	0	0	0	0	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	0	0	0	0	1	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	0	0	1	0	0	1	0	0	0
M06	6	1	0	0	1	0	0	0	0	0
M06	7	0	0	0	1	0	0	0	1	0
M06	8	0	0	0	1	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	0	0
M06	10B	0	0	0	0	0	0	1	0	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	0	1	0
M07	1	1	0	1	0	0	0	0	1	0
M07	2	0	1	0	0	0	0	0	0	0
M07	3	0	1	0	1	0	0	0	0	0
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	1	0
M07	6	0	0	0	0	1	0	1	0	0
M07	7	0	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	0	0	0
M07	9	0	0	0	1	0	1	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	0	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	1	0	0	0	0	0	1	0

Table 12 Q-matrix refined by the stepwise method: Singapore

Block	Item	Attribute								
		1	2	3	4	5	6	7	8	9
M01	1	0	1	0	0	0	0	0	0	0
M01	2	0	0	0	0	0	0	0	1	0
M01	3	1	0	1	0	0	0	0	1	0
M01	4a	1	0	1	0	0	0	0	1	0
M01	4b	1	0	1	0	0	0	0	1	0
M01	4c	1	0	1	0	0	0	0	1	0
M01	5	0	0	1	1	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0
M01	8	0	1	0	0	0	0	0	1	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	0	1	0	0	0	0	0	0
M02	4	0	1	0	0	0	0	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	0	0	0	1	0	1	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	1	0	0	1	0	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	1	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	1	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	1	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	1	0	1	0	0	0	0	0	0
M03	10	1	0	0	1	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	1	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0
M03	17	0	1	0	0	0	0	0	1	0
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	0	1	0	0	0	0
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0

Table 12 (continued)

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M05	6	1	1	0	1	0	0	0	0	0
M05	7	0	1	0	0	1	0	0	0	0
M05	8	0	0	0	0	0	1	0	0	0
M05	9	0	0	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	0	0	0	0	0	1	0	0
M05	12	0	0	0	0	0	0	1	0	0
M05	13	0	1	0	0	0	0	0	0	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	0	0	0
M06	2	0	1	0	0	0	0	0	0	0
M06	3	0	0	0	0	1	0	0	1	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	0	0	0	1	0	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	0	0	1	0	0	0	0	0	0
M06	6	1	1	0	1	0	0	0	0	0
M06	7	1	0	0	1	0	0	0	0	0
M06	8	0	0	0	1	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	0	0
M06	10B	0	0	0	0	0	0	1	0	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	0	1	0
M07	1	1	1	0	0	0	0	0	0	0
M07	2	0	1	0	0	0	0	0	0	0
M07	3	0	1	0	1	0	0	0	0	0
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	0	0
M07	6	0	0	0	0	1	0	1	0	0
M07	7	1	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	1	0	0
M07	9	1	0	0	1	0	0	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	0	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	0	0	0	0	0	0	1	0

Table 13 Q-matrix refined by the stepwise method: Australia

Block	Item	Attribute								
		1	2	3	4	5	6	7	8	9
M01	1	1	0	0	0	0	0	0	0	0
M01	2	0	0	1	0	0	0	0	0	0
M01	3	1	0	1	0	0	0	0	0	0
M01	4a	1	0	1	0	0	0	0	0	0
M01	4b	1	0	1	0	0	0	0	1	0
M01	4c	1	0	1	0	0	0	0	0	0
M01	5	0	0	1	1	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0
M01	8	0	1	0	0	0	0	0	1	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	0	0	0	0	0	0	0	0
M02	4	0	1	0	0	0	1	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	0	0	0	1	0	1	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	1	0	0	1	0	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	1	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	0	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	1	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	0	0	1	0	0	0	0	0	0
M03	10	1	0	0	1	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	1	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0
M03	17	0	1	0	0	0	0	0	1	0
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	1	0	0	0	0	0
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0

Table 13 (continued)

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M05	6	1	0	0	1	0	0	0	0	0
M05	7	0	1	0	0	1	0	0	0	0
M05	8	1	0	0	0	0	1	0	0	0
M05	9	0	0	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	0	0	0	0	0	1	0	0
M05	12	0	0	0	1	0	0	1	0	0
M05	13	0	1	0	0	0	0	0	0	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	1	0	0
M06	2	0	1	0	0	0	0	0	0	1
M06	3	1	0	0	0	0	0	0	1	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	0	0	0	1	0	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	1	0	1	1	0	0	0	0	0
M06	6	1	0	0	1	0	0	0	0	0
M06	7	0	1	0	1	0	0	0	0	0
M06	8	0	0	0	0	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	0	0
M06	10B	0	0	0	0	0	0	1	0	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	1	1	0
M07	1	1	0	1	0	0	0	0	1	0
M07	2	0	1	0	1	0	0	0	0	0
M07	3	0	1	0	1	0	0	0	0	0
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	0	0
M07	6	0	0	0	0	1	0	1	0	0
M07	7	0	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	1	0	0
M07	9	1	0	0	1	0	0	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	0	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	0	0	0	0	0	0	1	0

Table 14 Q-matrix refined by the stepwise method: Tunisia

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M01	1	1	0	0	0	0	0	0	0	0
M01	2	0	0	1	0	0	0	0	0	0
M01	3	1	0	1	0	0	0	0	0	0
M01	4a	0	0	1	0	0	0	0	0	0
M01	4b	0	0	1	0	0	0	0	0	0
M01	4c	0	0	1	1	0	0	0	0	0
M01	5	0	0	1	0	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0
M01	8	0	1	0	0	0	0	0	1	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	0	0	0	0	0	0	0	0
M02	4	0	1	0	0	0	0	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	0	0	0	0	0	1	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	1	0	0	0	0	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	1	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	0	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	0	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	1	0	1	0	0	0	0	0	0
M03	10	1	0	0	0	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	1	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0
M03	17	0	1	0	0	0	0	0	0	0
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	0	0	0	0	0	0
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0

Table 14 (continued)

Block	Item	Attribute								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M05	6	1	0	0	1	0	0	0	0	0
M05	7	0	0	0	0	1	0	0	0	0
M05	8	1	0	0	0	0	1	0	0	0
M05	9	0	0	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	0	0	0	0	0	1	0	0
M05	12	0	0	0	0	0	0	1	0	0
M05	13	0	1	0	0	0	0	0	0	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	0	0	0
M06	2	0	1	0	0	0	0	0	0	0
M06	3	1	0	0	0	0	0	0	1	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	1	0	0	0	0	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	0	0	1	1	0	0	0	0	0
M06	6	1	0	0	1	0	0	0	0	0
M06	7	0	0	0	1	0	0	0	0	0
M06	8	0	0	0	0	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	0	0
M06	10B	0	0	0	0	0	0	1	1	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	0	1	0
M07	1	1	0	0	0	0	0	0	1	0
M07	2	0	1	0	0	0	0	0	0	0
M07	3	0	1	0	1	0	0	0	0	0
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	0	0
M07	6	0	0	0	0	1	0	1	0	0
M07	7	1	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	1	0	0
M07	9	1	0	0	0	0	0	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	0	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	0	0	0	0	0	0	1	0

Acknowledgements
Not applicable.

Author contributions
JD and CC contributed equally to this work as first authors. They prepared the initial draft of the analyses and manuscript. JYP and WVN supervised the study and revised and commented on the draft. All authors read and approved the final manuscript.

Author's information
Jolien Delafontaine is a PhD student at the Faculty of Psychology and Educational Sciences, Parenting and Special education Unit of the KU Leuven. Her doctoral research focuses on effective teaching for students with special educational needs (SEN).
Changsheng Chen is a PhD student at the Faculty of Psychology and Educational Sciences, and the imec research group itec of the KU Leuven. His doctoral research focuses on learning analytics.
Jung Yeon Park is an assistant professor of quantitative research methods at George Mason University. Her research focuses on cognitive diagnosis models, large-scale educational assessments, and learning analytics.
Wim Van den Noortgate is a professor of statistics at the Faculty of Psychology and Educational Sciences, and the imec research group itec of the KU Leuven. His major interests include learning analytics and meta-analysis.

Funding
The authors received no specific funding for this work.

Availability of data and materials
The datasets analyzed during the current study are TIMSS 2011 at https://timssandpirls.bc.edu/timss2011/international-database.html.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 10 February 2022 Accepted: 24 October 2022
Published online: 21 November 2022

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705
Baker, F. B. (2001). *The basics of item response theory*. Retrieved from http:///ericae.net.irt/baker.
Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: Comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education: Principles, Policy & Practice, 12*(2), 167–181. https://doi.org/10.1080/09695940500143852
Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2–14. https://doi.org/10.1111/emip.12020
Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*(4), 277–293. https://doi.org/10.1177/0146621616686021
Chiu, C. Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618. https://doi.org/10.1177/0146621613488436
Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal Mathematics, Science & Technology Education*. https://doi.org/10.12973/eurasia.2015.1421a
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. https://doi.org/10.1111/j.1745-3984.2008.00069.x
de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130. https://doi.org/10.3102/1076998607309474
de la Torre, J. (2011). The Generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7
de la Torre, J., & Chiu, C. Y. (2016). General method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8
Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In: *International Conference on Artificial Intelligence in Education* (pp. 441–450). Berlin: Springer.
Groß, J., Robitzsch, A., & George, A. C. (2016). Cognitive diagnosis models for baseline testing of educational standards in math. *Journal of Applied Statistics, 43*(1), 229–243. https://doi.org/10.1080/02664763.2014.1000841
Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.
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. https://doi.org/10.1080/13803611.2010.523294

Jia, B., Zhu, Z., & Gao, H. (2021). International Ccomparative study of statistics learning trajectories based on PISA data on Cognitive Diagnostic Models. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.657858
Johnson, M.S., Lee, Y.S., Park, J.Y., Zhang, Z., & Sachdeva, R. (2013). *Comparing attribute distribution across countries: Application to TIMSS 2007 mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
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. https://doi.org/10.1177/01466210122032064
Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing, 14*(1), 49–72. https://doi.org/10.1080/15305058.2013.835728
Köhn, H. F., & Chiu, C. Y. (2016). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*(1), 112–132. https://doi.org/10.1007/s11336-016-9536-7
Köhn, H. F., & Chiu, C. Y. (2018). How to build a complete Q-matrix for a cognitively diagnostic test. *Journal of Classification, 35*(2), 273–299. https://doi.org/10.1007/s00357-018-9255-0
Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287–296. https://doi.org/10.2307/1391878
Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.
Liu, J. (2015). On the consistency of Q-matrix estimation: A commentary. *Psychometrika, 82*(2), 523–527. https://doi.org/10.1007/s11336-015-9487-4
Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357–383. https://doi.org/10.1177/0013164416685599
Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement, 43*(5), 402–414. https://doi.org/10.1177/0146621618798664
Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*(1), 3–26. https://doi.org/10.3102/1076998615621293
Ma, W., & de la Torre, J. (2020a). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14
Ma, W., & de la Torre, J. (2020b). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163. https://doi.org/10.1111/bmsp.12156
Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory Models. *Measurement: Interdisciplinary Research & Perspective, 11*(3), 71–101. Doi: https://doi.org/10.1080/15366367.2013.831680
Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(3), 333–356. https://doi.org/10.1080/10705511.2011.581993
Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075
Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. International Association for the Evaluation of Educational Achievement (IEA). Amsterdam: IEA Secretariat.
Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement, 79*(4), 727–753. https://doi.org/10.1177/0013164418822700
Park, J. Y., Lee, Y. S., & Johnson, M. S. (2017). An efficient standard error estimator of the DINA model parameters when analyzing clustered data. *International Journal of Quantitative Research in Education, 4*(1/2), 244–264. https://doi.org/10.1504/ijqre.2017.10007548
R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation, 20*(1), 11. https://doi.org/10.7275/5g6f-ak15
Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics, 4*(1), 87–94. https://doi.org/10.2307/1391390
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136
Sedat, ŞE. N., & Arican, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Eğitimde Ve Psikolojide Ölçme Ve Değerlendirme Dergisi, 6*(2), 238–253. https://doi.org/10.21031/epod.65266
Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104
Tatsuoka, K. K. (1984). Analysis of errors in fraction addition and subtraction problems. Final Report. Retrieved from University of Illinois, Computer-Based Education Research Lab website: https://files.eric.ed.gov/fulltext/ED257665.pdf.
Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. https://doi.org/10.1037/1082-989x.11.3.287
Terzi, R., & de la Torre, J. (2018). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education, 5*(2), 248–262. https://doi.org/10.21449/ijate.407193

von Davier, M., & Lee, Y. S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer Publishing.
Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement, 42*(6), 446–459. https://doi.org/10.1177/0146621617752991
Wu, X., Wu, R., Chang, H. H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.02230
Zheng, Y., Chiu, C.-Y., & Douglas, J. (2019). NPCD: Nonparametric methods for cognitive diagnosis; R Package Version 1.0–11. https://CRAN.R-project.org/package=NPCD

Publisher's Note