


RESEARCH

Open Access



Investigating item complexity as a source of cross-national DIF in TIMSS math and science

Qi Huang^{1*} , Daniel M. Bolt¹ and Weicong Lyu¹

*Correspondence:
qhuang85@wisc.edu

¹ Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI, USA

Abstract

Background: Large scale international assessments depend on invariance of measurement across countries. An important consideration when observing cross-national differential item functioning (DIF) is whether the DIF actually reflects a source of bias, or might instead be a methodological artifact reflecting item response theory (IRT) model misspecification. Determining the validity of the source of DIF has implications for how it is handled in practice.

Method: We demonstrate a form of sensitivity analysis that can point to model misspecification induced by item complexity as a possible cause of DIF, and show how such a cause of DIF might be accommodated through attempts to generalize the IRT model for the studied item(s) in psychometrically and psychologically plausible ways.

Results: In both simulated illustrations and empirical data from TIMSS 2011 and TIMSS 2019 4th and 8th Grade Math and Science, we have found that using a form of proposed IRT model generalization can substantially reduce DIF when IRT model misspecification is at least a partial cause of the observed DIF.

Conclusions: By demonstrating item complexity as a possible valid source of DIF and showing the effectiveness of the proposed approach, we recommend additional attention toward model generalizations as a means of addressing and/or understanding DIF.

Keywords: Large scale assessment, Cross-national DIF, Item complexity, ICC asymmetry, IRT model misspecification, TIMSS

Introduction

Large scale international assessments depend on the invariance of measurement across countries (Rutkowski & Svetina, 2014). Differential item functioning (DIF) analyses can be a significant component of such evaluations. Despite a multitude of techniques for the statistical detection of DIF in cross-national assessments, the ability to explain individual occurrences of DIF often remains a challenge (El Masri & Andrich, 2020; Zumbo, 2007). The issue is important, as it is generally appreciated that DIF can occur for valid or invalid reasons, and determining the validity of the source of DIF has implications for how it is handled in practice. Items showing substantial DIF for invalid reasons may be best removed from the assessment (e.g., Martin et al., 2020), while items showing DIF

for presumably valid reasons might instead be accommodated through models of partial invariance (Robitzsch & Lüdtke, 2020) for example, or possibly even ignored.

A somewhat complicating factor in DIF analyses for international assessments is that the groups (countries) being compared often have substantially different latent mean proficiency levels (Martin et al., 2020; OECD, 2017; Tijmstra et al., 2020). Under such conditions, it is known that even seemingly negligible amounts of model misspecification in an item can contribute to the statistical detection of DIF when no DIF is actually present (Bolt, 2002). By definition, DIF implies an observed difference in the relationship between latent proficiency and item score, commonly represented as an item characteristic curve (ICC), across groups. Effectively, the presence of model misfit leads to different parameters yielding better fit to the item response data depending on the proficiency distributions of the respondents from the given country. Countries having students concentrated at different locations along the latent proficiency continuum can be expected to yield different item parameter estimates for the misspecified model, implying false observation of DIF. An important applied feature of item response theory models, namely the invariance of item parameters, is likely lost when the item response model no longer fits the data (Shepard et al., 1984).

In recent years it has become further appreciated how IRT model misspecification is likely present at some level even for validly functioning items. Increasingly researchers have questioned the appropriateness of models like the 2PL or 3PL as accurate representations of what are often widely varying psychological response processes associated with different items. In this paper, we focus on studying potential misspecifications of the 2PL model in large scale international assessments, and investigating how one specific type of generalization of the 2PL, the logistic positive exponent (LPE; Samejima, 2000) model can reduce DIF. While more general models are commonly used for scoring in TIMSS, models such as the Rasch and 2PL are frequently used as a basis for DIF studies (e.g., Foy et al., 2016; Valdivia Medinaceli et al., 2023). The LPE model is attractive as a generalization of the 2PL because of its close connection to aspects of psychological response process, especially for items in subject areas like math and science. Samejima (2000), for example, noted how different forms of asymmetry can be expected in ICCs in the presence of either disjunctive or conjunctive interactions among multiple latent component response processes. Items that can be solved through the use of different strategies (e.g., a problem-solving OR a guessing strategy) often display negative ICC asymmetry, while items requiring multiple cognitive steps to solve (e.g., a complex mathematics word problem) often display positive ICC asymmetry, for example. Lee and Bolt (2018) observed that the magnitude of asymmetry seems also to be influenced by characteristics of the component processes (i.e., variability in discrimination across components) such that significant asymmetries in the ICCs are possible even when only two component processes are involved. Empirical studies that allow ICCs to possess asymmetry consistently show superior fit (Bazán et al., 2006; Bolfarine & Bazán, 2010; Lee & Bolt, 2018; Molenaar, 2015; Shim et al., 2022), with Lee & Bolt (2018) observing statistically detectable ICC asymmetries among TIMSS 8th grade Math items. Thus, the potential for model misspecification to contribute to DIF on TIMSS assessment seems high, and systematic attempts to study model misspecification as a possible source of DIF seem worthwhile.

Bolt & Liao (2021) demonstrated how the presence of unmodeled ICC asymmetry can lead to artificial DIF detection when using either “nonparametric” approaches, such as the Mantel–Haenszel or Standardization methods, or parametric approaches (e.g., through use of the Rasch or 2PL models) to evaluate DIF. Their results follow from the theoretical observation that ICC asymmetry interacts with difficulty such that expected scores conditional upon test sum scores will produce systematic differences in the empirical ICCs across groups that differ substantially in proficiency. Thus problems in evaluating DIF seemingly arise across broad classes of commonly used methods of DIF detection if the true ICCs assume certain functional forms.

One of the appealing aspects of contemporary IRT model estimation tools is their increased capacity to fit more complex (and yet psychologically plausible) models to item response data. Although most investigators adopt a common IRT model to apply across all test items, there may often be reasons to allow modifications of the IRT model on an item-by-item basis where needed. In this paper, we show how such a strategy may be justified and desirable in accommodating certain DIF items when the DIF may be attributed to model misspecification. We focus in particular on DIF seen across groups (countries) of widely varying proficiency distributions, where the effects of model misspecification are anticipated to be most profound. Using an index for DIF based on Oshima et al. (2015)’s multigroup generalization of Raju et al. (1995) Noncompensatory Differential Item Functioning (NCDIF) approach, we can examine how DIF may in many instances decline substantially when the item’s functional form is allowed to change in valid and plausible ways. For example, it is conceivable that an item showing substantial DIF under the 2PL will show substantially reduced DIF when fit with a model assuming positive ICC asymmetry (e.g., the logistic positive exponent model (LPE) with a specified exponent parameter of 15). In such instances, we might in turn then inspect the item more closely to confirm whether this characterization of the item as being of high complexity seems appropriate.

Besides lending greater insight into the underlying causes of DIF, we argue that an approach of item-specific model generalization may be preferred to specifying models of partial invariance, a common alternative strategy, in accommodating items that initially display DIF. The reason is that the proposed model generalization approach still allows the affected items to contribute to cross-national comparisons of proficiency, while models of partial measurement invariance only allow the affected item to function in support of within-country comparisons of proficiency. Where item and test specifications play a large role in test development, the loss of items of a particular type or category for purpose of country comparisons likely undermines the meaningfulness of the ultimate country comparisons. We comment further on this issue later in the paper.

The remainder of the paper is organized as follows. First we present the NCDIF approach (Oshima et al., 2015; Raju et al., 1995) used to quantify DIF, followed by a description of the models used to study DIF. Next, we provide simulation illustrations of the issue described above, and demonstrate how item-level model generalization of a form accommodating ICC asymmetry provides a mechanism for reducing DIF when model misspecification is at least a partial cause of observed DIF. Then, we examine application of the technique to real data item responses from TIMSS 2011 and TIMSS 2019 Math and Science at both the 4th grade and 8th grade levels. For illustration

purposes, we only focus on the countries of highest and lowest mean proficiency levels in the empirical analysis. In addition to providing an overall summary of the findings, we highlight as illustrations several example items for which the strategy appears effective in accommodating the observation of DIF. Finally, we revisit the comment above concerning the use of the proposed model generalization versus one of partial measurement invariance in allowing the original DIF item to reassume a role in cross-national comparisons.

Quantifying two-group and multiple group DIF

As will be described shortly, our simulation analyses initially consider DIF in a multi-group setting where each group represents a different country, and distinct item parameters are estimated for the studied item in each country. This reflects an ideal condition in which a sufficient number of respondents per country allow for estimation of IRT models at the country level. However, in many large-scale international assessments, such as TIMSS, there are often not enough respondents to estimate IRT model-based item parameters at the country level, necessitating a slightly different approach in which we collapse respondents from low mean proficiency and high mean proficiency countries into two groups. This approach still allows us to demonstrate the phenomenon of interest, namely the tendency to see the emergence of DIF due to model misspecification when the groups compared vary substantially in their proficiency distributions. Consequently, our simulation considers both multigroup and two-group DIF applications.

In both applications, we adopt quantifications of DIF based on the NCDIF strategy, formally presented by Raju et al. (1995) in the two-group case (see also Wainer, 1993) and generalized by Oshima et al. (2015) to the multi-group case. As there is often not a well-defined “reference group” in many multigroup applications (such as in many international assessment applications), we slightly alter the NCDIF approach as presented in Oshima et al. (2015) for our multigroup application so as to characterize the variability seen in ICCs around the mean ICC across all countries. Note that since this mean ICC is defined from the model-based probabilities of multiple countries, in the current application it likely does not conform to any particular IRT model. We can however use a discrete approximation approach to quantify such an index. Let $P_{ig}(\theta)$ denote the ICC for item i in country g . In the current analyses, we define $P_{i*}(\theta)$ to be the average of the ICCs across countries:

$$P_{i*}(\theta) = \frac{\sum_{g=1}^G P_{ig}(\theta)}{G},$$

although in theory $P_{i*}(\theta)$ could alternatively be defined from an IRT model with equal item parameters across countries. Our resulting NCDIF* index is then defined as

$$NCDIF_i^* = \frac{\sum_{g=1}^G \int (P_{ig}(\theta) - P_{i*}(\theta))^2 f_g(\theta) d\theta}{G}$$

where $f_g(\theta)$ represents the density of proficiency for country g , and $P_{ig}(\theta)$ denotes the model-based item response function for country g once all countries are placed on the same latent metric. Consistent with Raju et al. (1995) and Oshima et al. (2015),

in calculating $NCDIF_i^*$ we assume all items with the exception of item i have the same parameters (no DIF) across groups so that the parameters of item i can be viewed against a common metric.

To connect to the real data analyses, we also consider a two-group application by collapsing students from high mean proficiency countries into one group (here denoted as a reference group) and those from all low proficiency countries into a second group (here denoted as a focal group), as the full multi-country IRT analysis applied in the simulation does not support model fitting at the individual country level due to the small sample sizes of some countries. This approach still preserves our intent of demonstrating how the added flexibility afforded by allowing the ICCs to be asymmetric can relieve much of the DIF otherwise seen in the data, and also allows us to connect simulation results to the empirical analysis described in the next section. In the two-group condition the NCDIF statistic in which both ICCs above are defined using an IRT model can then be calculated in its more traditional way as

$$NCDIF_i = \int (P_{iR}(\theta) - P_{iF}(\theta))^2 f(\theta) d\theta$$

where $f(\theta)$ is the proficiency density of the focal group.

To evaluate the meaningfulness of NCDIF, we adopt the guideline provided by Oshima et al. (2015) and Wright & Oshima (2015), where 0.003 and 0.008 are the thresholds for “moderate” and “large” amounts of DIF, respectively. (Note that for ease of presentation, in the ensuing quantification of NCDIF we multiply these values by 100, making 0.3 and 0.8 the corresponding cutoffs.) While an item parameter replication (IPR) method was also applied in Oshima et al. (2015) to evaluate the statistical significance of NCDIF, it is not considered in this paper, as the statistical significance of DIF is less central to our illustrations than its quantification.

It is important to comment on other approaches to quantifying DIF and explain our reasoning for our use of $NCDIF^*$ in this context. Seemingly the most popular recent approach applied in large scale international assessment for evaluating DIF at the country-level is the root-mean squared deviation (RSMD; von Davier, 2017) index, which for a given country g , evaluates the empirical conditional probabilities observed for students in the country g about the expected (model-based) conditional probability, typically defined from all countries. The RMSD index for a given item i in country g can be written as

$$RMSD_{ig} = \sqrt{\int (P_{ig,o}(\theta) - P_{i,e}(\theta))^2 f_g(\theta) d\theta}$$

but where discrete approximation of $P_{ig,o}(\theta)$ is based on empirically observed (rather than model-based) conditional probabilities for item i in group g . $P_{i,e}(\theta)$ is obtained by using the model-based estimated item parameters. As has been noted elsewhere, this approximation adds to the NCDIF index elements of model misfit in group g (which may or may not also be present in the estimated ICC for other countries) as well as error related to sampling variability. The separation of these elements may not always be important in evaluating DIF; however, as our goal was one of demonstrating the degree (proportional reduction) in model-based DIF that can be observed when generalizing the model (i.e., the ability to observe a 0 on the index when model-based DIF is

successfully removed), we used NCDIF as basis for our quantifications, but do provide some example illustrations of the corresponding RMSD values in example items.

IRT models: the 2PL and LPE generalizations

There are now various psychometric models that have been introduced that can flexibly accommodate the presence of asymmetry (Falk & Cai, 2016; Molenaar, 2015; Robitzsch, 2022; Samejima, 2000). As indicated above, different forms of asymmetry can be anticipated in ICCs due to the presence of either disjunctive or conjunctive interactions among multiple latent component response processes. Among the various models accommodating asymmetric ICCs, the LPE has been most closely discussed in relation to psychological response processes of these kinds. These features are also embodied in the concept of item complexity (Samejima, 2000), a concept often considered when developing items in content areas like math and science. For many math and science items, problem solving strategies can be understood in relation to multiple cognitive steps to solutions (high complexity); other items can be solved using multiple strategies and/or provide opportunity for proficiency-based guessing to play a role (low complexity). Under such conditions, it is anticipated that the 2PL, assuming a common form of symmetry for all items, has a high potential to produce misspecification for at least some of the items. However, the LPE model, due to its discussion in relation to such aspects of response process as well as its ease of implementation using the *mirt* (Chalmers, 2012) package, is considered as a promising alternative and is used in the current study.

The LPE model accommodates different levels of item complexity by generalizing ordinary IRT models (Rasch, 2PL, 3PL) with an extra exponent parameter, ξ ($\xi > 0$). Following Samejima (2000), the larger the exponent parameter ξ , the more complex the item. Note that $\xi = 1$ implies symmetry; ICCs display positive asymmetry when $\xi > 1$, and negative asymmetry when $0 < \xi < 1$. The LPE is equivalent to the 2PL when $\xi = 1$.

Statistically, the item response function of 2PL can be statistically written as:

$$P(Y_{ij} = 1|\theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

while for the LPE model it is:

$$P(Y_{ij} = 1|\theta_i) = \left[\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \right]^{\xi_j}$$

Despite the anticipated plausibility of the LPE, it has become increasingly appreciated that difficulties in estimating its parameters will often make it an empirically intractable model to freely estimate across all items. When estimation is attempted, the item parameter ξ is frequently found to be confounded with the item b parameter; in addition, the desirable invariance properties of the item parameters under the 2PL model are likely no longer present when the underlying model is LPE. However, as we seek to illustrate in this paper, such limitations do not preclude meaningful use of the model in the context of multigroup analyses where some subset of the items might be constrained as $\xi = 1$, (i.e., treated as 2PL) while other items are constrained at different levels of ξ so as to minimize DIF, especially where such alternative levels of ξ may be psychologically

defensible. This makes use of the model plausible in the context of multiple-group analysis. As person and country-level scoring can readily accommodate the presence of alternative values of ξ for individual items, it may well be of value to specify models that allow for different ξ values, even if found in an exploratory way, rather than either omitting the item from overall scoring (e.g., by dropping the item from analysis) or omitting the item from between-country comparisons (e.g., by fitting a model of partial invariance).

Appendix 3 provides an illustration of R code used to specify analyses in which all items but the studied item are fitted as 2PL items and constrained to have equal parameters across groups, while for the studied item both the a and b are allowed to vary across groups. For the current analyses, the ξ parameter of the studied item is considered at various levels, and the resulting NCDIF is quantified. Of particular interest is the degree to which NCDIF declines when the ξ parameter is set at values other than 1, reflecting the 2PL.

Simulation illustrations of model misspecification-induced DIF and DIF resolution through item specific model generalization

In this section we illustrate via simulation how asymmetry-related model misspecification yields DIF for groups that vary substantially in mean proficiency level. We in turn show how generalizing the IRT model for the suspect item has the potential to minimize the quantified DIF. Respondents from twenty groups (e.g., countries, $k = 1, \dots, 20$) are simulated having mean proficiency levels uniformly distributed from $\mu_k = -1.8$ to 1.8. As these analyses are designed to be illustrative, we minimize effects of estimation error by simulating a sample size of 10,000 respondents for each group having latent proficiencies $\theta_{ik} \sim Normal(\mu_k, 1)$. For each respondent we generate responses for $j = 1, \dots, 20$ items, assuming 19 2PL items and 1 LPE item. For the 2PL items, $b_j \sim Uniform(-2, 2)$, and $a_j \sim Lognormal(0, 0.25)$. For the LPE items, we consider two generating conditions: in one, the studied item has positive ICC asymmetry and the other the studied item has negative ICC asymmetry. For the positive asymmetric item, we let $b = -3, a = 0.9, \xi = 15$, while the negative asymmetric item has $b = 2, a = 1.5, \xi = 0.4$. Both sets of item parameters construct plausible ICCs (see the red curves labeled as “true ICC” in Fig. 1). Importantly, in all cases we generate the same item parameters across countries, meaning that no actual DIF is present. Thus, any DIF observed would be an artifact of using the “wrong model” (combined with should be minimal amounts of model estimation error to our use of large respondent samples).

While various estimation algorithms for multigroup IRT analyses exist, in the current paper we use the *mirt* (Chalmers, 2012) routine in R. An appealing feature of *mirt* in the current setting is its capacity to specify alternative functional forms for the ICC, one of which can be an LPE model with a specified asymmetry parameter ξ . Using multigroup data, this feature provides the opportunity to initially (1) quantify cross-national DIF for the studied item using NCDIF* under the baseline model (e.g., 2PL) and then (2) quantify the reduction in NCDIF* when the studied item is fit as an LPE with a specified asymmetry parameter ξ . We approach (1) using a measure of DIF variability quantified from the estimated ICCs for each group. To estimate NCDIF* for the studied item, we specify a multigroup IRT model in which all item parameters except for the studied item are assumed equal across countries, and the proficiency distribution for each country

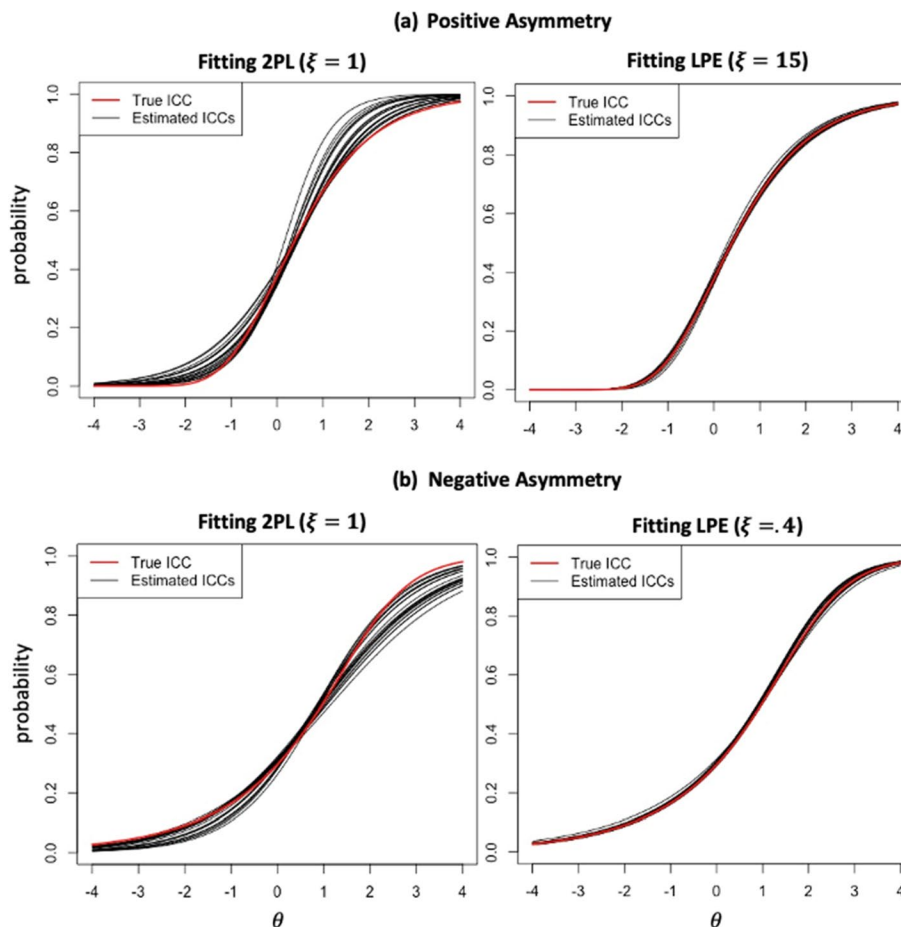


Fig. 1 Studied Item ICCs Across Groups When Fitting Studied Item as 2PL and LPE, Simulation Illustration

is assumed normal but with estimated mean. NCDIF* is then quantified using a discrete approximation technique that quantifies the variance in the estimated ICCs across countries using a sequence of proficiency values (from -6 to 6 in increments of 0.01) weighted by the pooled proficiency distribution across countries. In addition, as the most substantial DIF usually occurs when the mean proficiencies of the two countries differ substantially, we also calculate the NCDIF index for only the highest mean proficiency country and the lowest mean proficiency country. We approach (2) in the form of a sensitivity analysis, by considering different potential values for the asymmetry parameter and then observing how NCDIF* (and the two-country NCDIF) changes. For the two-group analysis, we again specify a multigroup model as before constraining all items but the studied item to have equal 2PL parameters across groups, while the studied item is now fit as an LPE with one of six levels specified for the asymmetry parameter ($\xi = 0.2, 0.4, 0.7, 8, 15$ and 20) while estimating the corresponding a and b parameters. Although attending to multiple levels of asymmetry in the LPE seems unnecessary in the simulation, it naturally proves necessary in real data analysis where the true ξ is unknown. This sensitivity analysis approach is similar to that reported by Bolt & Liao (2021) in examining how the correlation between DIF and difficulty changes when allowing items to assume varying asymmetry parameters.

Table 1 DIF quantifications of studied item when fitting the 2PL and LPE with different specified ξ levels, simulation illustration

Positive asymmetry generating condition				Negative asymmetry generating condition			
True ξ	Specified ξ	NCDIF	NCDIF*	True ξ	Specified ξ	NCDIF	NCDIF*
15	1 (2PL)	0.817	0.076	0.4	1 (2PL)	0.519	0.052
15	8	0.017	0.004	0.4	0.2	0.1551	0.029
15	15	0.007	0.0043	0.4	0.4	0.0142	0.004
15	20	0.005	0.0043	0.4	0.7	0.0261	0.023

NCDIF* denotes a multiple-group generalization of the NCDIF index (Oshima et al., 2015)

Figure 1(a) and (b) illustrate the results seen for the simulated LPE items under the multigroup analysis. Each curve reflects the ICC for the studied item within a particular country when fitting the misspecified 2PL (left figure) and the correctly specified LPE (right figure). Clearly the variability of the group ICCs substantially decreases when the correctly specified LPE is fitted. Corresponding results are seen in the NCDIF* indices of Table 1, which also shows the NCDIF index when comparing only the groups of highest and lowest mean proficiency levels. In particular, for the positive asymmetry condition, large DIF (> 0.8) is observed when fitting the 2PL, but it is greatly reduced to a negligible level (0.007) when the LPE with the correct ξ is fitted. Regarding the negatively asymmetric item, the moderate level of NCDIF (0.519) seen under the 2PL shrinks to 0.014 when fitting the LPE with the correctly specified ξ . Although the presence of estimation error still keeps the variability of ICCs slightly above 0 even when the model is correctly specified, the variability is significantly reduced when the correct level of asymmetry is applied, and is reduced even for conditions in which at least the direction (if not precise level) of asymmetry is correct.

Real data analyses: TIMSS 2011 and TIMSS 2019 Math and Science at grades 4 and 8

To examine whether the strategy above demonstrates any meaningful reduction in DIF for actual items on TIMSS, we conducted separate analyses for TIMSS 2011 and TIMSS 2019, Math and Science 4th and 8th Grade Assessments by booklet. It was not our belief that all, or even a majority of items would necessarily show diminished DIF under this approach, but rather that some would, and that we might further justify the LPE specification that minimizes DIF if the associated asymmetry level could be justified from inspection of the item. To the extent that this general approach can be applied to also consider other plausible forms of model misspecification, our primary goal is simply to see whether the model generalization approach has the potential to meaningfully reduce the observation of DIF.

As mentioned earlier, the real data structure differs from the simulation in that there are often too few respondents to an item from any one country to fit the IRT model at the country level and obtain item parameter estimates with sufficient precision. Thus, in the current analysis we collapse countries with the lowest and highest mean proficiencies into two groups—a low proficiency group and a high proficiency group—making the analysis a two-group analysis. The sample size for each of the 6 assessments (i.e., TIMSS

Table 2 Proportions of total items showing more than a 50% reduction in NCDIF Index, empirical TIMSS analyses

	Total # of items	Math		Science	
		Positive asymmetry	Negative asymmetry	Positive asymmetry	Negative asymmetry
2019					
Grade 4 paper	261	0.129	0.154	0.053	0.124
Grade 4 electronic	261	0.117	0.147	0.074	0.088
Grade 8 paper	325	0.201	0.201	0.106	0.093
Grade 8 electronic	325	0.198	0.169	0.090	0.123
2011					
Grade 4	347	0.233	0.264	0.143	0.170
Grade 8	434	0.249	0.214	0.186	0.172

2019 4th grade paper-based and electronic, 8th grade paper-based and electronic, and TIMSS 2011 paper-based forms for both 4th and 8th grades) is approximately equal to 50% of all examinees who took this assessment, which varies between 1500 to 7000 as the number of total examinees taking each assessment is different. To satisfy the intention of including approximately 50% of all examinees, we choose the number of countries being collapsed in each assessment accordingly. For instance, in TIMSS 2011 4th grade Math, 15 countries with highest mean proficiencies and 15 countries with lowest mean proficiencies were collapsed into the two groups respectively. This also allows us to calculate NCDIF index as in the simulation.

We conducted analyses across both paper-based and electronic forms of the TIMSS 2019 assessments. Similar to what has been done in the simulation study, for each booklet, we fit both an ordinary multigroup 2PL model and the multigroup LPE with several fixed asymmetry parameter levels, with each item taking a turn as the studied item. Studied items were evaluated for DIF using asymmetry parameters of 0.2, 0.4, 0.7, 3, 8, and 15 (an exponent value of 1 corresponds to the 2PL). As the asymmetry parameter for the studied item is consistently fixed (as opposed to estimated), we view our approach as a form of sensitivity analysis in which we can observe how quantifications of DIF change at different levels of assumed asymmetry. Based on the item parameters obtained for the studied item in each group, we calculate the corresponding NCDIF value to see if re-specifying the 2PL by adding an exponent parameter would help reduce DIF. For items scored using partial credit (initially scored 0, 1, 2), we transform the score as binary, such that original scores of 0, 1 were coded incorrect and 2 correct.

While the likelihoods are essentially indistinguishable when fitting the 2PL and LPE, the quantification of DIF can be substantially different. Table 2 reports the proportion of items on each assessment type for which NCDIF was found to reduce 50% or more through use of the LPE with a specified ξ parameter as opposed to when fitting the 2PL. As seen in Table 2, up to half of the Math items have more than a 50% reduction in NCDIF (depending on the form) when fitting LPE with an asymmetry parameter other than 1. Around half of these improvements occur when specifying positive ICC asymmetry, while the other half when specifying negative ICC asymmetry. For Science, the numbers are slightly more modest, with between 16 and 35% showing DIF

reduction of at least 50% through asymmetry, with again nearly half of the improvements occurring through positive asymmetry, half through negative asymmetry. A more detailed breakdown of the DIF reductions can be found in Appendix 1, where the items are classified into three categories based on the level of DIF—large, moderate, and small—they initially show when the 2PL is applied. Compared to the large DIF category, it seems clear that a greater proportion of moderate and small DIF items show a larger-than-50% reduction when fitting LPE with asymmetry parameter other than 1. Nevertheless, there are improvements seen in all three categories.

We consider some examples of TIMSS 2011 items for which allowing asymmetry seems particularly useful in terms of DIF reduction. Appendix 2 displays the actual items used for these examples (all of which have been released). As seen from these example items, the DIF observed under the 2PL appears related as much (if not more) to variability in discrimination as difficulty. This is as expected. ICCs that are positively asymmetric produce reduced discrimination in the ICCs of the higher proficiency countries; negatively asymmetric items just the reverse.

For 4th grade Math, items M031016 and M031218 provide examples where allowing positive ICC asymmetry and negative ICC asymmetry, respectively, reduced NCDIF by more than 50% (note that RMSD also reduces by approximately 40%, at least for one group, although as noted earlier, RMSD is sensitive to more than just DIF). Figure 2 displays the ICCs for these items (in each case allowing for distinct a and b parameters across groups) when the 2PL (i.e., $\xi = 1$) is specified versus when $\xi = 15$ (for M031016) and $\xi = 0.2$ (for M031218). By comparing Fig. 2(a) and (c), as well as Fig. 2(b) and (d), it is clear that fitting a multi-group LPE with a specified ξ to the studied items can substantially reduce variability in ICCs, further suggesting that use of the LPE can be an effective alternative to 2PL in reducing cross-national DIF. Furthermore, inspection of the content of the items in Appendix 2 suggests each of these items displays the hallmark features of items suspected to show these corresponding patterns of asymmetry, providing a psychological justification of the application of LPE model. For item M031016, a correct answer consists of correctly specifying the complete list of numbers satisfying the condition; any one miss implies an incorrect answer. Such multicomponential items are anticipated to show positive asymmetry. As a multiple-choice item, M031218 is suspected to possess negative ICC asymmetry due to the presence of multiple solution strategies. As shown in Lee & Bolt (2018) and Bolt & Liao (2022), the negative asymmetry becomes particularly pronounced when the guessing process has a weak positive relationship to the proficiency.

We find similar types of examples in 8th grade Math. Figure 3 shows corresponding pairs of ICCs for items M052206 and M032419 (item content shown in Appendix 2), which similarly show conditions where allowing positive ICC asymmetry and negative ICC asymmetry, respectively, substantially reduce the observed DIF, and for similar plausible reasons to those observed for the Grade 4 Math items.

As seen in the table above, the reductions of DIF appear to occur for larger numbers of Math items than Science items. This may be a reflection of a tendency for Math items to show greater psychological response process variability, at least of a kind that manifests as ICC asymmetry. Examples of Science items that displayed similar reductions in DIF can nevertheless be found in Appendix 2.

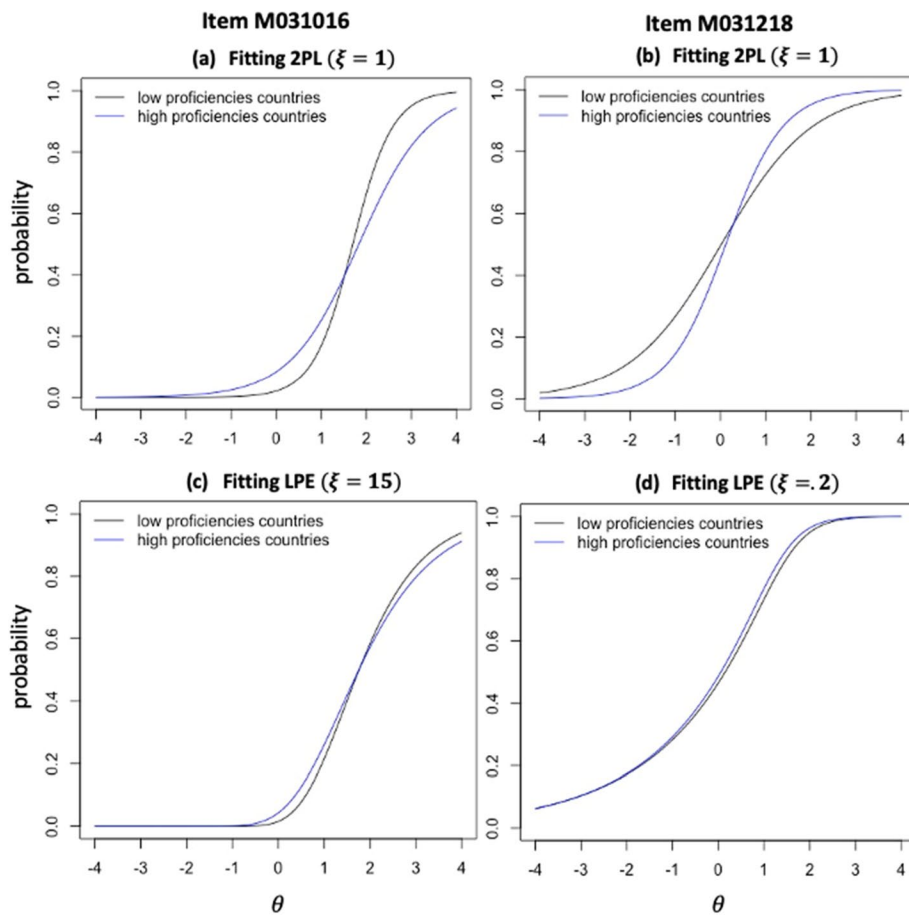


Fig. 2 ICCs When Fitting 2PL and LPE to the Studied Item, 4th Grade Math, TIMSS 2011. For Item M031016, the NCDIF index decreases from 0.964 to 0.094 when fitting with 2PL versus LPE having $\xi = 15$, while the changes in RMSD are from 0.023 to 0.018 for the low proficiency country group, and from 0.036 to 0.017 for the high proficiency country group. For Item M031218, the NCDIF index reduces from 0.464 to 0.056 when fitting 2PL versus LPE having $\xi = 0.2$, while the changes in RMSD, are from 0.041 to 0.025 for the low proficiency country group, and from 0.019 to 0.018 for the high proficiency country group

Model re-specification versus models of partial measurement invariance in accommodating DIF

As noted earlier, a common approach to accommodating DIF involves fitting models of partial invariance, whereby the parameters for the DIF item are allowed to vary for the affected countries, as opposed to maintaining the same parameters across countries. It might be questioned what the advantage is of a strategy that seeks to accommodate DIF items by generalizing the psychometric model (as illustrated in this paper) as opposed to one that proceeds with a model of partial measurement invariance to accommodate the DIF. With partial measurement invariance, particularly when 2PL item parameters are freed for the affected items across countries, these freed items no longer contribute to between-country comparisons of proficiency. Specifically, if we assume that all items that maintain the same parameters in the partial measurement invariance model yield an ordering of proficiency means across countries of $\mu_1, \mu_2, \mu_3, \dots, \mu_G$, it should be appreciated that the performances on DIF

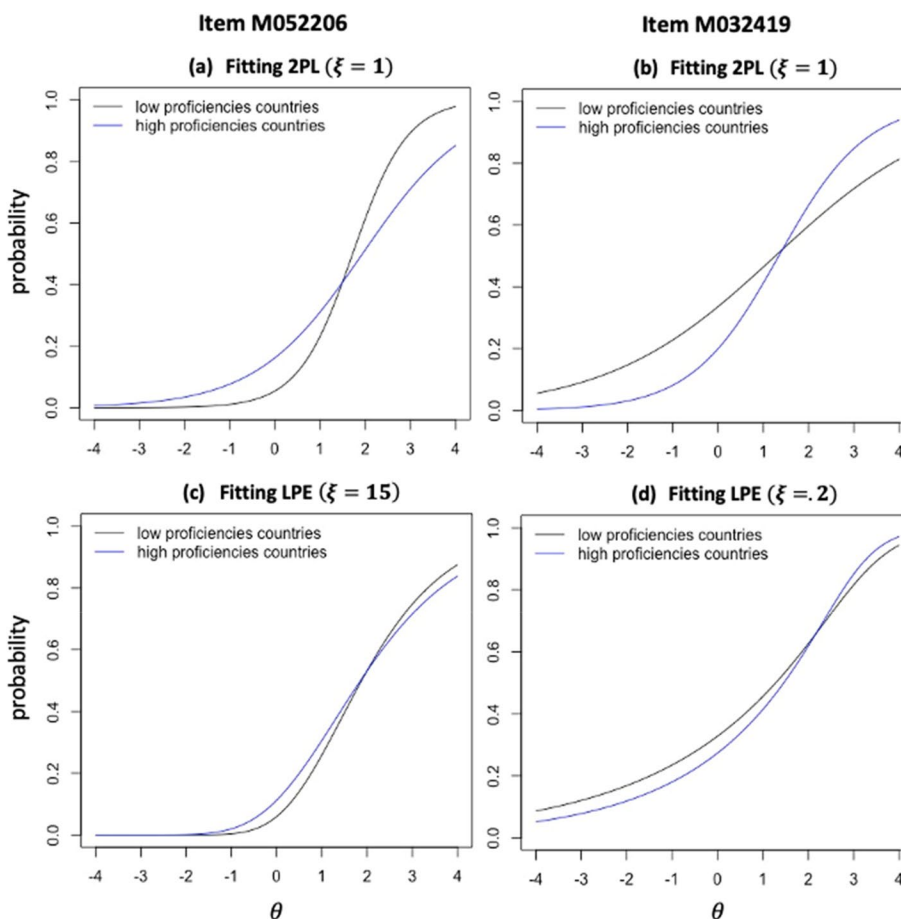


Fig. 3 ICCs When Fitting 2PL and LPE to the Studied Item, 8th Grade Math, TIMSS 2011. For Item M052206, the NCDIF index decreases from 1.447 to 0.110 when fitting 2PL versus LPE having $\xi = 15$, while the changes in RMSD are from 0.046 to 0.034 for the low proficiency country group, and from 0.041 to .021 for the high proficiency country group. For Item M032419, the NCDIF index reduces from 1.075 to 0.421 when fitting 2PL versus LPE having $\xi = 0.2$, while the changes in RMSD, are from 0.038 to 0.015 for the low proficiency country group, and from 0.048 to 0.021 for the high proficiency country group

items no longer contribute to the variability seen in the country means. Thus while the DIF item may still contribute to how the proficiency distinguishes respondents within the country, it will not contribute any additional information to the estimates of mean country proficiencies. To the extent that country level proficiency comparisons are often a primary interest in international large scale assessments, this might be viewed as a significant loss. However, identifying an alternative model for the item, such as an LPE with appropriately chosen ξ parameter that successfully resolves or minimizes DIF, would allow the item to contribute to estimates of country level proficiencies, while still distinguishing respondent proficiencies within country. We would suggest that even where this is not statistically consequential, it may still be perceived as beneficial from a measurement perspective, as the between-country comparisons can be more confidently understood in relation to a more consistent representation of the content areas represented on the assessment. A more formal exploration of such consequences is left to future research.

Discussion and conclusion

Our results suggest a strong likelihood that model misspecification contributes to the statistical emergence of cross-national DIF among at least some of TIMSS Math and Science items. The result occurs primarily for groups (countries) that are separated substantially in terms of mean proficiency level. Items that display sizeable NCDIF under the 2PL frequently show substantial declines in NCDIF when the fitted model permits asymmetry in the ICC. When this is the case, the pattern of DIF under the 2PL generally conforms to what is expected in the presence of positive or negative asymmetry in the ICC for the item. High complexity (positive ICC items) tends to be less discriminating for high proficiency (relative to low proficiency) groups, while just the reverse occurs for low complexity items. We are also able to successfully connect these psychometric properties to the items themselves. Specifically, the positive asymmetry items were commonly ones that were open-ended (as opposed to multiple-choice) and demanded multiple cognitive steps to reach a correct answer; in some cases, the correct answer itself had multiple components. In addition, those items scored as partial credit, which generally entail multiple steps and were in the present analysis scored as correct only when achieving full credit, also frequently had DIF under the 2PL that was substantially reduced when a positive exponent parameter was specified.

One advantage to trying to address the emergence of country level DIF under this approach is that it allows the originally DIF items to still contribute to cross-country comparisons in the measured proficiency. A partial measurement invariance approach similarly accommodates DIF items, but by allowing the item parameters to assume different values for different countries, the item no longer contributes to comparisons of proficiency levels across countries, only the estimation of the proficiency within country. In this respect, the proposed approach may be superior. We acknowledge this advantage may well only be of intrinsic value; it may not yield any demonstrable effect when applied in contexts where most items lack DIF. However, whenever a significant DIF is present under traditional IRT models (i.e., the discarded items in TIMSS), we suggest the researchers closely inspect the item to see if such DIF can be reduced by applying the approach shown in this paper.

To the extent that our sensitivity analysis approach encourages investigators to reflect on psychological response process in relation to model generalization, we believe our approach also has value in understanding sources of DIF, increasingly becoming a goal of DIF investigations (Zumbo, 2007). While it has been noted that model generalization by itself is unlikely to make DIF completely disappear (von Davier & Bezirhan, 2023), observing substantial reductions in DIF under model generalization, even with some misfit remaining, seemingly has the potential to help understand why items show DIF—a frequent challenge in DIF studies (e.g., El Masri & Andrich, 2020). Given that understanding sources of DIF can inform subsequent item and test development practices, we think the value of such practice should not be overlooked.

There are many additional directions that could be pursued in this work. Of course our use of the LPE represents only one form of alternative model that could be considered. A similar approach could be attempted with other more general IRT models. There may also be more effective ways of automating the process of identifying the exponent parameter of the LPE that potentially minimizes DIF. The current approach

was designed only to demonstrate the effect; practical use of the methodology would naturally benefit from automated procedures with greater efficiency. In this respect it is important to also appreciate that our illustrations only considered a small range of discrete ξ values. Values larger than 15 on the positive end or lower than 0.2 on the negative end are entirely plausible when considering that the interacting psychological response processes may vary substantially in discrimination (Bolt & Liao, 2022; Lee & Bolt, 2018). Appendix 4 shows for one example item how an even greater reduction of DIF may be achieved by even more extreme values of ξ than were considered in this study.

We should also acknowledge that the NCDIF index used in this paper is just one way of quantifying DIF. Other indices, such as the root means square deviation (RMSD; Tijmstra et al., 2020) statistic or the differential response functioning (DRF; Chalmers, 2018) statistic, can also be taken into consideration in the future. In addition, our analyses are also made inefficient by the general inability to estimate all parameters of the LPE simultaneously. The limitation is mitigated to some extent because the LPE is only applied to one item in the current application, with all other items being specified as 2PL. Nevertheless, estimation problems with the LPE have been documented (Lee, 2015). While our analyses were restricted to sensitivity analyses in which the asymmetry parameter was consistently specified at chosen values, alternative models for asymmetry might facilitate easier estimation of the asymmetry parameter for a single item (with all other item constrained). Finally, alternatives to use of the 2PL as the base model could also be considered, including the 3PL, for example.

We hope that our simple proof of concept might stimulate more attention toward model misspecification as a possible source of DIF that is observed in practice. We are of course not claiming that all DIF has this as its source; however, attention to this as a possibility might minimize some of the costs associated with removing items from group comparisons when such items ultimately do not need removal.

Appendices

Appendix 1

Table 3 is a breakdown of Table 2 in the paper according to the amount of DIF observed under the 2PL.

Table 3 Number (proportion) of items showing more than a 50% reduction in NCDIF Index for each category of DIF (i.e., large, moderate, small), averaged across booklets, empirical TIMSS analyses

	Science											
	Math								Science			
	Positive asymmetry		Negative asymmetry		Positive asymmetry		Negative asymmetry		Positive asymmetry		Negative asymmetry	
	Large	Mod.	Small	Large	Mod.	Small	Large	Mod.	Small	Large	Mod.	Small
2019												
G4P	0.3 (0.03)	1.0 (0.12)	2.5 (0.19)	0.5 (0.06)	0.9 (0.11)	3.0 (0.22)	0.1 (0.02)	0.5 (0.06)	0.9 (0.07)	0.1 (0.01)	1 (0.12)	2.6 (0.21)
G4E	0.3 (0.04)	1.6 (0.16)	2.2 (0.12)	0 (0.00)	1.2 (0.12)	3.8 (0.21)	0.2 (0.04)	0.6 (0.06)	1.6 (0.13)	0.4 (0.04)	0.9 (0.08)	1.8 (0.12)
G8P	2.8 (0.17)	2.4 (0.27)	1.5 (0.14)	2.6 (0.15)	2.2 (0.24)	1.8 (0.17)	1.2 (0.06)	1.4 (0.13)	1.5 (0.13)	0.6 (0.03)	0.8 (0.12)	2.1 (0.23)
G8E	2.0 (0.17)	2.5 (0.28)	3.7 (0.14)	1.2 (0.10)	1.9 (0.20)	4.1 (0.16)	0.4 (0.02)	0.9 (0.12)	2.9 (0.12)	0.4 (0.02)	0.9 (0.10)	4.4 (0.20)
2011												
G4	1.6 (0.18)	2.1 (0.32)	2.3 (0.22)	2.4 (0.28)	1.2 (0.19)	3.1 (0.28)	2.4 (0.13)	0.7 (0.15)	0.9 (0.18)	2.7 (0.14)	1.2 (0.31)	0.9 (0.16)
G8	3.1 (0.21)	2.6 (0.33)	1.9 (0.21)	3.7 (0.25)	1.4 (0.18)	1.5 (0.17)	3.9 (0.20)	1.4 (0.20)	1.1 (0.16)	2.4 (0.12)	1.9 (0.23)	1.6 (0.22)

Mod. stands for moderate level of DIF ($0.3 \leq \text{NCDIF} < 0.8$). G4 and G8 represent Grade 4 and Grade 8 respectively; the P and E following G4 and G8 refer to the paper-test and electronic-test formats. Note that in TIMSS 2011, the paper-test was the only available format

Appendix 2

This includes the content of TIMSS 2011 Math example and Science example items in which similar sources of asymmetry appear plausible and are seen to reduce NCDIF.

Example Items:

1. **Item M031016** (4th grade Math), displaying positive ICC asymmetry.

ID: M031016	Mathematics Grade 4	Block_Seq: M05_02
M031016	<p>Three thousand tickets for a basketball game are numbered 1 to 3000. People with ticket numbers ending with 112 receive a prize. Write down all the prize-winning numbers.</p>	Content Domain
	Prize-winning numbers: _____	Number
		Topic Area
		Whole Numbers
		Cognitive Domain
	Reasoning	Maximum Points
	1	Key
	See scoring guide	

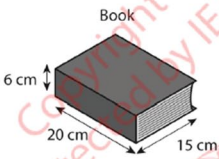
2. **Item M031218** (4th grade Math), displaying negative ICC asymmetry.

ID: M031218	Mathematics Grade 4	Block_Seq: M05_09
M031218	<p>Six hundred books have to be packed into boxes that hold 15 books each. Which of the following could be used to find the number of boxes needed?</p>	Content Domain
	(A) add 15 to 600	Number
	(B) subtract 15 from 600	Topic Area
	(C) multiply 600 by 15	Whole Numbers
	(D) divide 600 by 15	Cognitive Domain
	Applying	Maximum Points
	1	Key
	D	

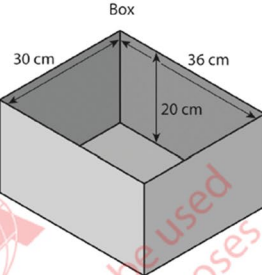
3. Item M052206 (8th grade Math), displaying positive ICC asymmetry.

ID: M052206	Mathematics Grade 8	Block_Seq: M02_12
-------------	---------------------	-------------------

Ryan is packing books into a rectangular box.
All the books are the same size.



Book



Box

What is the largest number of books that will fit inside the box?

Answer: _____

Content Domain
Geometry

Topic Area
Geometric Measurement

Cognitive Domain
Reasoning


Maximum Points
1

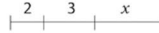
Key
See scoring guide


4. Item M032419 (8th grade Math), displaying negative ICC asymmetry.


ID: M032419	Mathematics Grade 8	Block_Seq: M05_04
-------------	---------------------	-------------------

Which of these could represent the expression $2x + 3x$?

(A) The length of this segment: 

(B) The length of this segment: 

(C) The area of this figure: 

(D) The area of this figure: 

Content Domain
Algebra

Topic Area
Algebraic Expressions

Cognitive Domain
Knowing

Maximum Points
1

Key
C

5. Item S032650Z (8th grade Science), displaying positive ICC asymmetry.

ID: S032650Z	Science Grade 8	Block_Seq: S07_12
--------------	-----------------	-------------------

Tamora is preparing to climb one of the highest mountains on Earth. She knows that the atmospheric conditions will change the higher up the mountain she climbs.

In the table below, write down two atmospheric conditions that will change as Tamora climbs the mountain. State what Tamora needs to bring in order to survive these two conditions at high elevations.

Change in Atmospheric Condition	What Tamora Needs to Bring
1.	
2.	

Content Domain
Earth Science

Topic Area
Earth's Structure and Physical Features

Cognitive Domain
Applying

Maximum Points
2

Key
See scoring guide

6. Item S041180 (4th grade Science), displaying negative ICC asymmetry.

ID: S041180	Science Grade 4	Block_Seq: S06_04
-------------	-----------------	-------------------

The diagram below shows a food chain.

```

graph LR
    A[green algae] --> B[krill]
    B --> C[fish]
    C --> D[seal]
    D --> E[killer whale]
            
```

green algae krill fish seal killer whale

Which predator-prey relationship is correct?

(A) fish (predator)-seal (prey)
 (B) green algae (predator)-krill (prey)
 (C) fish (predator)-krill (prey)
 (D) seal (predator)-killer whale (prey)

Content Domain
Life Science

Topic Area
Ecosystems

Cognitive Domain
Applying

Maximum Points
1

Key
C

In addition to the content of the Science example items, we also provide their ICC plots when fitting 2PL versus LPE (Fig. 4). Similar to Figs. 2 and 3, it is clear in Fig. 4 that fitting a multi-group LPE with a specified ξ to the studied items can substantially reduce variability in ICCs, further suggesting that use of the LPE can be an effective alternative to 2PL in reducing cross-national DIF.

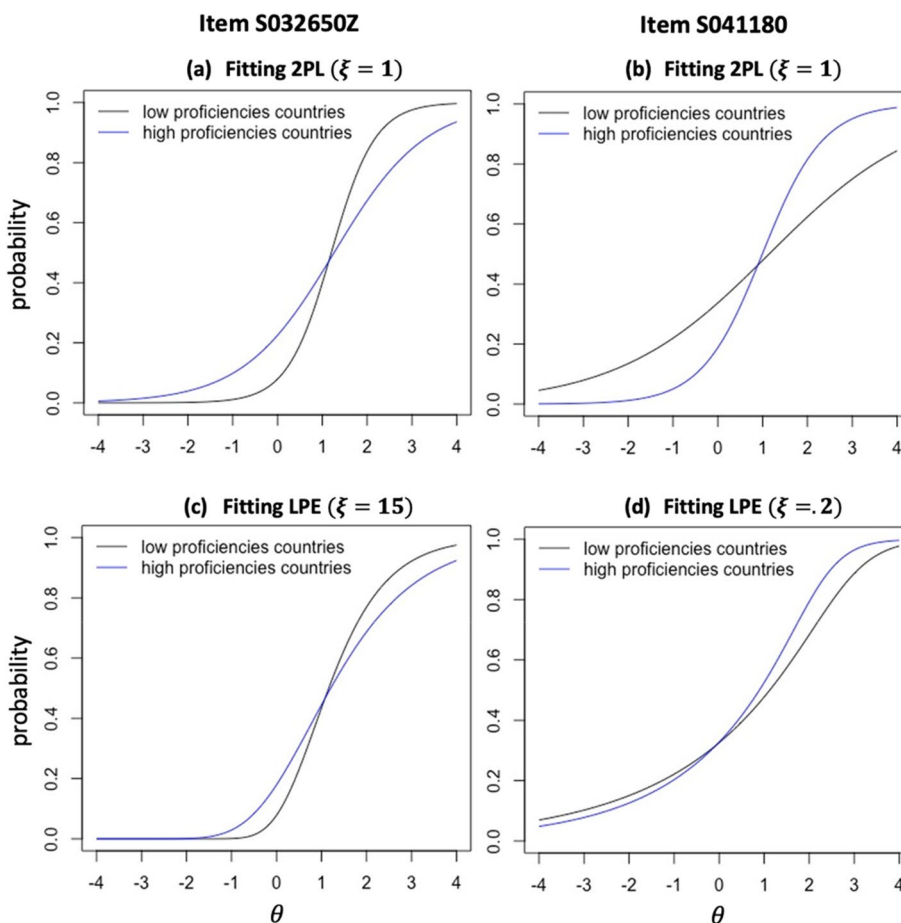


Fig. 4 ICCs when fitting 2PL and LPE to the studied item, TIMSS 2011 Science

Appendix 3

We provide examples of R code used to specify analyses where the studied item is fitted as LPE using a fixed value of ξ but estimated discrimination (a) and difficulty (b) parameters for each group, while all other items are fitted as 2PL items and constrained to have equal parameters across groups. This modeling can be achieved through the *mirt* R package (Chalmers, 2012). The analysis consists of three steps: (1) specify the LPE ξ parameter for the studied item, (2) specify priors for a and b parameters, (3) fit the multi-group IRT model and estimate item parameters. In the first step, we define the LPE model by specifying the level of exponent parameter (i.e., $\xi = 3$), while allowing a and b to be freely estimated. Note that $a = 1$ and $b = 0$ are just used as starting values, instead of fixed values. Next, we specify priors for a and b on the studied item so that the item parameter estimates are sensitive to the choice of ξ but largely noninformative. In the third step, we use “multipleGroup()”, a built-in function in *mirt*, to fit the multi-group model to the dataset. To simulate a condition in which only the studied item is fitted as LPE while all the other items are fitted as 2PL, we use the “itemtype=” argument. Only the studied item (which is item 14 in this case) is fit as LPE; all other items

are 2PL. To impose equality constraints on the discrimination and difficulty parameters across groups for the 2PL items, we specify the item names using the “invariance=” argument. Note that the studied item is excluded from the “invariance=” argument, as it is the only item allowed to have varying a and b across groups, reflecting its status as the item potentially displaying DIF.

1. First Step: define LPE function.

```
# define LPE
name <- 'LPE'
par <- c(a=1, b = 0)
expn <- 3 # adjust the exponent parameter for sensitivity analysis
estimate <- c(TRUE, TRUE)
P.LPE <- function(par,Theta, ncat){
  a = par[1]
  b = par[2]
  P1 = plogis(a * (Theta - b))^expn
  cbind(1-P1, P1)
}
x <- mirt::createItem(name, par=par, est=estimate, P.LPE)
```

2. Second Step: specify priors.

```
# specify priors
mod <- '
  F = 1-23
  START[focal] = (14,a,1)
  START[ref] = (14,a,1)
  PRIOR[focal] = (14, a, lnorm, 0, 0.5)
  PRIOR[ref] = (14, a, lnorm, 0, 0.5)

  PRIOR[focal] = (14, b, norm, 0, 5)
  PRIOR[ref] = (14, b, norm, 0, 5)
,
```

3. Third Step: fit model and estimate item parameters.

```
#fit model and estimate item parameters
itemnames <- colnames(response)
fit <-
  mirt::multipleGroup(
    data = response, model = mod, group = group,
    itemtype = c(rep("2PL", 13), "LPE", rep("2PL", 9)),
    customItems = list(LPE=x),
    invariance = c("free_mean", "free_var", itemnames[-14]),
    # only let the studied item (item 14) have varied a and b across groups
    technical = list(NCYCLES=5000))
```

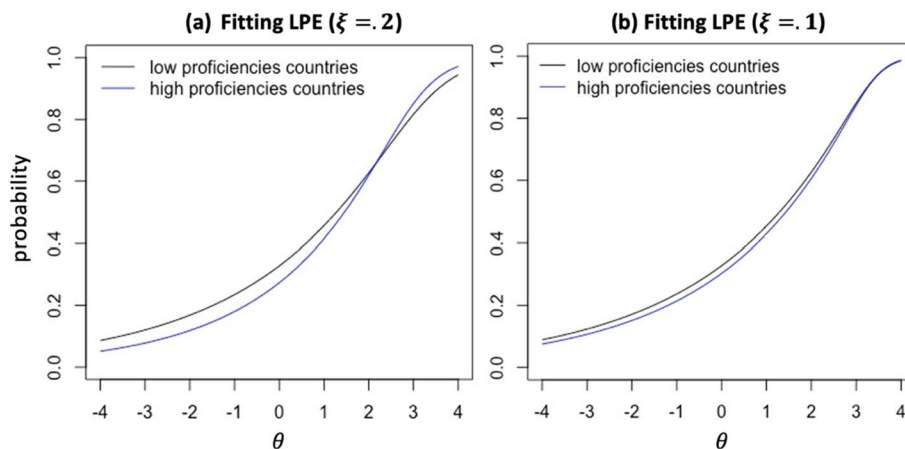


Fig. 5 ICCs when fitting LPE with extreme values of ξ to the studied item, item M032419, 8th Grade Math, TIMSS 2011

Appendix 4

Our paper only considers values of ξ less than or equal to 15 on the positive asymmetry end, and greater than or equal to 0.2 on the negative asymmetry end. More extreme ξ values are entirely possible, as may occur when the interacting response component processes vary substantially in discrimination (Bolt & Liao, 2021; Lee & Bolt, 2018). Here we show for one of the 8th grade Math items (M032419), which in the paper showed minimal DIF when $\xi = 0.2$ can be even further reduced by choosing an even lower ξ . Specifically, by fitting the LPE with $\xi = 0.1$ to this item, we observe an even lower NCDIF quantification of 0.026, a 73% reduction of the original NCDIF obtained in the paper when the LPE with $\xi = 0.2$ is fitted. The corresponding result in terms of the ICCs can be seen in Fig. 5. Although the difference between the two ICCs in Fig. 5(a) is already much smaller than when fitting 2PL (see Fig. 3 in the paper), Fig. 5(b) shows that the two ICCs become even closer when a more extreme value of ξ is considered.

Abbreviations

2PL	Two-parameter logistic
3PL	Three-parameter logistic
DIF	Differential Item Functioning
DRF	Differential Response Functioning
ICC	Item characteristic curve
IPR	Item parameter replication
IRT	Item Response Theory
LPE	Logistic positive exponent
NCDIF	Noncompensatory Differential Item Functioning
RMSD	Root means square deviation

Acknowledgements

Not applicable.

Author contributions

QH (acquisition, analysis, visualization and interpretation of simulation data, and TIMSS 2011 and 2019 Math data; conceptualization; investigation; methodology; writing—drafting, reviewing, and editing). DMB (conceptualization; investigation; methodology; project administration; supervision; validation; writing—drafting, reviewing, and editing). WL (acquisition, analysis, visualization and interpretation of TIMSS 2011 and 2019 Science data; Investigation; writing—reviewing and editing).

Funding

This research received no specific grant from any funding agency in the public, commercial, or organizations for the submitted work.

Availability of data and materials

The data and item content that support the findings of this study are openly available in the database of IEA TIMSS & PIRLS International Study Center at <https://timssandpirls.bc.edu/databases-landing.html>.

Declarations**Competing interests**

The authors declare that they have no competing interests.

Received: 8 March 2023 Accepted: 8 April 2024

Published online: 22 April 2024

References

- Bazán, J. L., Branco, M. D., & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis*, 1(4), 861–892.
- Bolfarine, H., & Bazán, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational and Behavioral Statistics*, 35(6), 693–713.
- Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 15(2), 113–141.
- Bolt, D. M., & Liao, X. (2021). On the positive correlation between DIF and difficulty: A new theory on the correlation as methodological artifact. *Journal of Educational Measurement*, 58(4), 465–491.
- Bolt, D. M., & Liao, X. (2022). Item complexity: A neglected psychometric feature of test items? *Psychometrika*, 87, 1195–1213.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika*, 83(3), 696–732.
- El Masri, Y. H., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Applied Measurement in Education*, 33(2), 174–188.
- Falk, C. F., & Cai, L. (2016). Semiparametric item response functions in the context of guessing. *Journal of Educational Measurement*, 53(2), 229–247.
- Foy, P., Martin, M. O., Mullis, I. V. S., Yin, L., Centurino, V. A. S., & Reynolds, K. A. (2016). Reviewing the TIMSS 2015 Achievement Item Statistics. In: M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 11.1–11.43). Retrieved from Boston College, TIMSS & PIRLS International Study Center website. <http://timss.bc.edu/publications/timss/2015-methods/chapter-11.html>
- Lee, S. (2015). *A comparison of methods for recovery of asymmetric item characteristic curves in item response theory*. [Unpublished masters thesis]. University of Wisconsin, Madison
- Lee, S., & Bolt, D. M. (2018). Asymmetric item characteristic curves and item complexity: Insights from simulation and real data analyses. *Psychometrika*, 83(2), 453–475.
- Martin M. O., von Davier M., Mullis I. V. (Eds.) (2020). *Methods and procedures: TIMSS 2019 technical report*. <https://timssandpirls.bc.edu/timss2019/methods/pdf/TIMSS-2019-MP-Technical-Report.pdf>
- Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, 80(3), 625–644.
- OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.
- Oshima, T. C., Wright, K., & White, N. (2015). Multiple-group noncompensatory differential item functioning in Raju's differential functioning of items and tests. *International Journal of Testing*, 15(3), 254–273.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias. *Applied Psychological Measurement*, 19(4), 353–368.
- Robitzsch, A. (2022). On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy*, 24(6), 760.
- Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, 62(2), 233–279.
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74, 31–57.
- Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, 65, 319–335.
- Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, 9, 93–128.
- Shim, H., Bonifay, W., & Wiedermann, W. (2022). Parsimonious asymmetric item response theory modeling with the complementary log-log link. *Behavior Research Methods*, 55, 200–219.
- Tijmstra, J., Bolsinova, M., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement*, 57(4), 566–583.
- Valdivia Medinaceli, M., Rutkowski, L., Svetina Valdivia, D., & Rutkowski, D. (2023). Effects of DIF in MST routing in ILSAs. *Large-Scale Assessments in Education*, 11(1), 22.
- von Davier, M. (2017). *Software for multidimensional discrete latent trait models*. Educational Testing Service.
- von Davier, M., & Bezirhan, U. (2023). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement*, 83(4), 740–765.

- Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Erlbaum.
- Wright, K., & Oshima, T. C. (2015). An effect size measure for Raju's differential item functioning for items and tests. *Educational and Psychological Measurement*, 75, 338–358.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.