Bundsgaard *Large-scale Assess Educ* (2019) 7:9
https://doi.org/10.1186/s40536-019-0077-2

Large-scale Assessments in Education

RESEARCH

Open Access

DIF as a pedagogical tool: analysis of item characteristics in ICILS to understand what students are struggling with

Jeppe Bundsgaard*

*Correspondence:
jebu@edu.au.dk
Danish School of Education,
Aarhus University,
Copenhagen, Denmark

Abstract

International large-scale assessments like international computer and information literacy study (ICILS) (Fraillon et al. in International Association for the Evaluation of Educational Achievement (IEA), 2015) provide important empirically-based knowledge through the proficiency scales, of what characterizes tasks at different difficulty levels, and what that says about students at different ability levels. In international comparisons, one of the threats against validity is country *differential item functioning* (DIF), also called item-by-country interaction. DIF is a measure of how much harder or easier an item is for a respondent of a given group as compared to respondents from other groups of equal ability. If students from one country find a specific item much harder or easier than students from other countries, it can impair the comparison of countries. Therefore, great efforts are directed towards analyzing for DIF and removing or changing items that show DIF. From another angle, however, this phenomenon can be seen not only as a threat to validity, but also as an insight into what distinguishes students from different countries, and possibly their education, on a content level, providing even more pedagogically useful information. Therefore, in this paper, the data from ICILS 2013 is re-analyzed to address the research question: *Which kinds of tasks do Danish, Norwegian, and German students find difficult and/or easy in comparison with students of equal ability from other countries participating in ICILS 2013?* The analyses show that Norwegian and Danish students find items related to computer literacy easier than their peers from other countries. On the other hand, Danish and, to a certain degree, Norwegian students find items related to information literacy more difficult. Opposed to this, German students do not find computer literacy easier, but they do seem to be comparably better at designing and laying out posters, web pages etc. This paper shows that essential results can be identified by comparing the distribution of difficulties of items in international large-scale assessments. This is a more constructive approach to the challenge of DIF, but it does not eliminate the serious threat to the validity of the comparison of countries.

Keywords: ICILS 2013, Proficiency scales, Differential item functioning, Computer and information literacy, Pedagogical use of tests

Introduction

International large-scale assessments like *Programme for International Student Assessment* (PISA), and the *International Association for the Evaluation of Educational Achievement* (IEA) studies *progress in international reading literacy study* (PIRLS) and *international computer and information literacy study* (ICILS) are most known for the so-called league tables, which provide information about the relative abilities of students across countries. But for teachers, teacher educators, and developers of teaching material, they can provide much more important empirically based knowledge of what characterizes tasks at different difficulty levels, and what that says about students at different ability levels: What can they be expected to do easily, what is their present zone of proximal development, and which tasks are they not yet able to perform? This knowledge is summed up in so-called described proficiency scales, which are developed on the basis of analyses of items of similar difficulty and detailed studies of tasks at a given difficulty interval (Fraillon et al. 2015; OECD 2014).

When constructing a measure, the constructor needs to assure that it measures the same way for different persons being measured. This is called *measurement invariance*. It means that the result of a test should not depend on anything else but the students' proficiency in the area the test is intended to measure. It should not matter what background the student comes from, or on the specific items used to test this specific student.

In international comparisons a number of factors can be a threat to measurement invariance. Typically in order to cover a broad excerpt of the construct, individual students receive only a subset of the items. If these items are not representative of the construct, the measure could be biased. By rotating the booklets or modules, test designers are able to minimize the potential consequences of this problem, but still the problem could persist and be difficult to identify if the total set of items were not covering the construct.

One of the serious threats against measurement invariance is country *differential item functioning* (DIF), also called item-by-country interaction. DIF is a measure of how much harder or easier an item is for a respondent of a given group as compared to respondents from other groups of equal ability (Holland and Wainer 1993). If students from one country find a specific item much harder or easier than students from other countries, it can impair the comparison of countries. Therefore, in international large-scale assessments great efforts are directed towards analyzing for DIF and removing or changing items that show DIF (e.g. Fraillon et al. 2015, p. 166ff.).

Nonetheless, DIF seems to be unavoidable in large-scale assessments like PISA and ICILS, and this has drawn heavy criticism, especially directed towards PISA (Kreiner and Christensen 2014). But from another angle, this phenomenon can be seen not only as a threat to validity, but also as an insight into what distinguishes students from different countries, and possibly their education, on a content level.

Research questions

In this paper, the data from ICILS 2013 (Fraillon et al. 2014) is re-analyzed to get a deeper understanding of what students from three North European countries, Denmark, Norway, and Germany, find difficult or easy as opposed to students from other countries.[1]

[1] The other countries/education systems participating in ICILS 2013, and included in the analysis were Australia, Chile, Croatia, the Czech Republic, Republic of Korea, Lithuania, Poland, the Russian Federation, the Slovak Republic, Slovenia, Thailand, and Turkey. The two Canadian provinces Newfoundland and Labrador and Ontario were also included. The City of Buenos Aires (Argentina), Denmark, Hong Kong SAR, the Netherlands, and Switzerland did not meet the sampling requirements, and were therefore not included in the international item estimation. In this paper, however, Denmark is included.

Thus, the research questions are as follows:

> Research question 1: Can challenging content areas be identified by grouping items with Differential Item Functioning?
> Research question 2: Which kinds of tasks do Danish, Norwegian, and German students find difficult and/or easy in comparison to students of equal ability from other countries participating in ICILS 2013?

International computer and information literacy study

ICILS measures computer and information literacy (CIL) according to the following definition: "an individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in society" (Fraillon et al. 2013, p. 17). ICILS divides CIL into two strands: (1) collecting and managing information, and (2) producing and exchanging information, each consisting of 3–4 aspects: 1.1 Knowing about and understanding computer use, 1.2 Accessing and evaluating information, 1.3 Managing information, 2.1 Transforming information, 2.2 Creating information, 2.3 Sharing information, and 2.4 Using information safely and securely (Fraillon et al. 2013, p. 18).

The construct is measured using an innovative computer-based test made up of four modules each consisting of an authentic storyline where students are asked, for example, to help organize an after-school activity. The items types range from multiple choice and short text answers to the production of web pages and posters using interactive software.

The data is analyzed using a uni-dimensional Rasch model. A two-dimensional model, relating to the two strands mentioned before, was also tested, but the two dimensions showed a very high correlation (0.96), and it was therefore decided to base the analysis on the more simple Rasch model (Fraillon et al. 2014, p. 73).

Differential item functioning in large-scale assessments

The concept of DIF was developed as an alternative to *item bias* to avoid an implicit (negative) evaluation of the consequences of an item functioning differently for a group of test takers (Angoff 1993). DIF is a statistical concept, while item bias is a social concept. In the context of international educational surveys, DIF is also referred to as item-by-country interaction.

DIF is generally seen as a problematic phenomenon, i.e. as an indicator of item bias, and the solution is therefore often to remove items that show DIF, or to treat the items as not-administered for the groups where they showed DIF, or to allow for country-specific item parameters. But sometimes items are important for the construct, and differences in different groups can be understandable and meaningful. For example, Hagquist and Andrich (2017) argue that stomach ache as an indicator of psychosomatic problems will have different interpretations in boys and girls, because girls can experience stomach ache in connection with their menstrual periods. They state that: "It turns out that in dealing with this DIF a critical issue is whether this potential source of the DIF should be considered relevant or irrelevant for the conceptualisation of psychosomatic problems and its applications" (Hagquist and Andrich 2017, p. 7). Therefore, they suggest not just to remove items, but also to resolve them by splitting them into two items, one for each group. This way, the information remain in the study, and the groups' different relations to the item is taken care of. This solution is also available to international educational surveys, but it would make it more difficult to explain the construct theoretically and to deduce proficiency scales from the data because they would be different in countries with different country-specific parameters.

A number of studies have discussed the consequences of DIF in international large-scale assessments. According to Kreiner and Christensen (2014), the "evidence against the Rasch model is overwhelming" in their secondary analyses of PISA 2006 data, and they argue that the DIF is seriously impairing the league tables.

Using an alternative statistical method based on a long-form market basket definition (Mislevy 1998), Zwitser et al. (2017) argue that they are able to take DIF into account, and at the same time provide final scores that are comparable between countries. In their analysis, model fit improves substantially if country-specific item parameters are included in the method, and the resulting league table is "nearly the same as the PISA league table that is based on an international calibration" (Zwitser et al. 2017, p. 225). They use this as evidence for the claim that "PISA methodology is quite robust against (non-)uniform DIF" (ibid.).

Most of the research on DIF in international large-scale assessments is looking for sources for DIF and finds it in differences in language, curriculum, or culture (Huang et al. 2016; Oliveri et al. 2013; Sandilands et al. 2013; Wu and Ercikan 2006), while other studies investigate gender DIF (Grover and Ercikan 2017; Innabi and Dodeen 2006; Le 2009; Punter et al. 2017).

Only one paper was found relating to DIF in ICILS 2013 (Punter et al. 2017). In this paper, the assessment data from ICILS 2013 was re-analyzed using a three-dimensional 2PL IRT model (the GPCM, generalized partial credit model), showing better fit than the Rasch model used in the international report (Fraillon et al. 2014). Correlations between these three dimensions ranged between 0.636 between dimension 1 and 3 for girls in Norway, and 0.982 between dimension 2 and 3 for girls in Slovenia.

The analysis of differences in boys and girls in these three dimensions showed that girls outperformed boys in most countries on the dimension called evaluating and reflecting on information, and even more so on the dimension called sharing and communicating information, while no significant gender differences were found in the dimension called applying technical functionality (Punter et al. 2017, p. 777). The authors argue that the DIF found in relation to gender, and resolved by implementing a three-dimensional solution, is an argument in favor of analyzing the ICILS data in three dimensions instead of the uni-dimensional solution chosen in the international report.

By far most of the research done in relation to DIF is concerned with improving test fairness, and has been for thousands of years (Holland and Wainer 1993, p. xiii). Therefore, the consequence of identifying DIF in items is usually to remove the item from the test or to resolve it by splitting it or marking it as not-administered for the groups showing DIF.

But as already pointed out by Angoff (1993, p. 21ff.), investigation of DIF can give interesting insights into the construct and into the groups of students taking the test etc. In some sense, each item in a test is a construct in itself, which tests the specific knowledge and/or skill that it asks about. Items in a test can typically be arranged into a number of groups of similar items, relating to a sub-area (aspect) of the construct. As noted, the items in ICILS are related to two strands, but in the international analysis, it was found that these strands were highly correlated, so the test can be considered as unidimensional (Fraillon et al. 2014, p. 73). In PIRLS and *trends in international mathematics and science study* (TIMSS) each of the three main constructs (science literacy, mathematical literacy, and reading literacy) are reported both as uni-dimensional scales and as multiple (three or four) sub-dimensions. The fact that countries are not positioned the same way in each league table of the sub-dimensions is an indication of differential item functioning in the main scale between items from the different subscales, which opens up for seeing "DIF as an interesting outcome" (Zwitser et al. 2017, p. 214). Thus, when defining a construct to be measured, one has to decide how broad it should be, and how much differential item functioning is acceptable between groups of items. This decision can be called DIF by design, which is also what is used in the analysis in this paper. DIF could be an indication of the instrument measuring more than one construct, but if the constructs are closely correlated, and conceptually connected, they might work adequately statistically for the majority of the groups of students. Finding DIF in items for a particular group, e.g., for students from a specific country, can therefore be seen both as an indication of multi-dimensionality of the construct and as a potentially interesting and important characteristic of this group, be it special skills or lack of knowledge.

Methods and instruments

The student responses found in the dataset from the *international computer and information literacy study* (ICILS) 2013 (Fraillon et al. 2014) are re-analyzed using the Rasch model (Rasch 1960). The Rasch model separates the item difficulties and the student abilities, making it possible to talk about item difficulties independently of the students taking the test. The Rasch model gives the probability of a correct response (a score of 1, rather than 0) to an item, i, with a difficulty (δ_i) depending on the ability (θ_p) of the respondent, p. When a respondent has the same ability as the difficulty of an item, she has a 50 percent probability of answering correctly. In case of items with more categories than 0 and 1, a partial credit version of the Rasch model can be used (Andersen 1977; Andrich 1978; Masters 1982). The partial credit model can be written as follows:

$$P_{pik} = \frac{e^{\sum_{j=0}^{k}(\theta_p - \delta_{ij})}}{\sum_{n=0}^{m_i} e^{\sum_{j=0}^{n}(\theta_p - \delta_{ij})}}, \quad k = 0, 1, \ldots m_i, \tag{1}$$

where P_{pik} is the probability of getting from category $k-1$ to k, and m_i+1 is the number of categories in item i.

Under the Rasch model, DIF can be described by the formula $P(X = 1|\theta, G) \neq P(X = 1|\theta)$, i.e. the probability of responding correctly to an item is different for members and non-members of the group G. This can be integrated into the Rasch model:

$$P_{pik} = \frac{e^{\sum_{j=0}^{k}(\theta_p - \delta_{ij} + \gamma_i G_p)}}{\sum\limits_{n=0}^{m_i} e^{\sum_{j=0}^{n}(\theta_p - \delta_{ij} + \gamma_i G_p)}}, \quad k = 0, 1, \ldots, m_i, \tag{2}$$

where G_p is 1 if the person is a member of the group p, and otherwise 0.

Given that item difficulties are estimated based on empirical data, they cannot be expected to be exactly the same for different groups. Therefore, a threshold for acceptable differences has to be set. Longford et al. (1993, p. 175) have reproduced a table developed by N.S. Petersen from the Educational Testing Service in 1987. In this table, Petersen differentiates between three categories of DIF based on Mantel and Haenszel's differential item functioning (MH D-DIF): A, B, and C. DIF in category A is so low that it is in no need of attention. In category B, the level of DIF calls for consideration, and "if there is a choice among otherwise equivalent items, select the item with the smallest absolute value of MH D-DIF" (ibid.). Items with a DIF in category C should only be included in a test if it is "essential to meet specifications", and should be documented and brought before an independent review panel.

Based on the *educational testing service* (ETS) DIF classification rules presented and expanded in Longford et al. (1993), Paek and Wilson (2011, p. 1028) calculate the threshold values as they would look in a Rasch framework:

A if $|\gamma| \leq 0.426$ or if $H_0 : \gamma = 0$ is not rejected below 0.05 level
B if $0.426 < |\gamma| < 0.638$ and if $H_0 : \gamma = 0$ is rejected below 0.05 level
C if $0.638 \leq |\gamma|$ and if $H_0 : \gamma = 0$ is rejected below 0.05 level

where A is considered a negligible DIF, B a medium DIF, and C a large DIF. In the ICILS DIF analyses that follow, the standard errors are well below 0.025 for all items. Thus in all cases, the null hypothesis will be rejected for γ above 0.426.

In ICILS, the international report selects 0.3 logits (described as "approximately about one-third of a standard deviation" of the distribution of students (Fraillon et al. 2015, p. 164) as the threshold for considerable DIF, which means that the difference between two groups would be 0.6 logits.

Results

The published[2] dataset from ICILS 2013 was used. Items were re-coded and deleted or excluded from the scaling for individual countries in accordance with the decisions in the technical report (Fraillon et al. 2015, pp. 171, 264ff.). This study intends to understand which content knowledge can be gained from items showing DIF, and, therefore, the items that were removed from the dataset prior to the final international estimation for the international report were kept for the countries of interest in this study. The reason for removal in the international study might very well have been DIF (but only one (A10C for Germany under the MH D-DIF Level C criteria) of the removed items was actually showing DIF in the analyses of the present study) (Additional file 1).

[2] https://www.iea.nl/our-data.

Table 1 Key parameters from the Rasch analyses

	# of items	# of cases	Deviance	EAP-reliability	# ICILS DIF	# DIF B	# DIF C
International	62	8000	298,222.10	0.90			
Norway	62	7500 + 500	297,839.48	0.90	9	6	14
Germany	61	7500 + 500	297,838.64	0.90	7	6	14
Denmark	62	7500 + 500	297,659.42	0.90	9	9	18

Number of cases is 8000 in all analyses. In the DIF-analyses 500 students from the country under investigation and 7500 students from other countries are included

The *test analysis modules* (TAM) package in R (Robitzsch et al. 2017) was used for the analyses,[3] which were carried out under conditions as close to the ICILS international study as possible, including the use of weights to sample a group of 500 students from each country (250 students from each of the two participating Canadian provinces). The model used was the partial credit model ("item + item * step"), and the estimation was done using the Marginal Maximum Likelihood algorithm, with the mean of the item difficulties constrained to 0. To make sure that the analysis was comparable to the international ICILS analysis, item difficulties from the estimation were compared to the item difficulties reported in the ICILS technical report (Fraillon et al. 2015, p. 171). One item (A10E) showed a rather large discrepancy (around 0.5 logits) from the ICILS estimations due to different response distributions in the samples.

In the ICILS international study, only countries that met the IEA sampling requirements were included in the estimation of the item difficulties. Because Denmark is one of the countries of interest in this study, it was included in the following analyses. In order to ensure comparability and soundness of the analyses, a comparison was made of an analysis of only the countries that met the sampling requirements with an analysis including 500 students from Denmark and the rest of the countries. Only minor differences were noted in the item difficulties in these two analyses.

The analyses of DIF were carried out individually for each country using the R formula $\sim item * step + country * item$ which is equivalent to the ConQuest (Wu et al. 2007) parametrization $\sim item + step + item * step + item + country + country * item.$ In the context of marginal maximum likelihood estimation, the analysis can take group differences in ability into account when estimating item parameters. This is done by allowing each group (in this case country and gender) to have their own population parameters.[4]

The standard settings of the TAM function tam.mml.mfr were used, except for fac.oldxsi set to 0.4 to ensure convergence. For comparison, an analysis of the same dataset was carried out without the country interaction. The results of these analyses are presented in Table 1.

As can be seen from the deviances, the analyses taking DIF into account do describe the data better for all countries in this study.

In order to test for homogeneity of the DIF, all expected score curves were plotted so the curves for the country under investigation could be compared visually to the curves

[3] The R code is available as an appendix to this paper.

[4] In the international estimation a group variable called "Windows" was included to take into account if a Windows computer was used by the test taker. This variable was not available in the public dataset.

Table 2 Easier and harder items for Danish students

Easier	Harder
Aspect 1.1: Knowing about and understanding computer use	*Aspect 1.1: Knowing about and understanding computer use*
Navigate to a URL given as plain text (− 0.23)	Open a link in a new browser tab (0.35)
Open a file of a specified file type (− 0.36)	*Aspect 1.2: Accessing and evaluating information*
Save a presentation with a new file name (− 0.23)	Select relevant images in a presentation (0.26)
Switch applications to an internet browser from the taskbar (− 0.47)	Presents accurate information (0.51)
Aspect 2.1: Transforming information	*Aspect 1.3: Managing information*
Excludes irrelevant information in a poster (− 0.32)	Use a flowchart template to design the navigational flow of a website (0.26)
Use software to crop an image (− 0.43)	*Aspect 2.1: Transforming information*
Adapt information for an audience (− 0.54)	Create a balanced layout of a webpage page (0.22)
Convert a description of directions into a visual route on a map (− 0.22)	*Aspect 2.2: Creating information*
Aspect 2.4: Using information securely and safely	Layout images in a presentation (0.23)
Identify that an email does not originate from the purported sender (− 0.42)	Create a balanced layout of text and images for an information sheet (0.3)
	Aspect 2.4: Using information securely and safely
	Identify information that is risky to include on a public profile (0.63)
	Identify that a link's URL does not match the URL displayed in the link text. (0.22)

Numbers in parenthesis are the DIF values

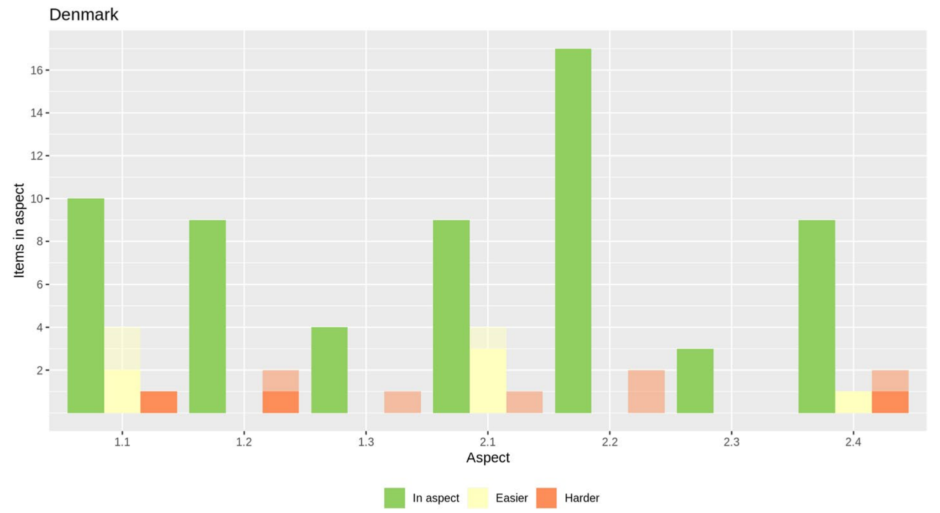

Fig. 1 Items showing DIF for Danish students. Green bars show how many items there are in the aspect in total. Yellow bars show the number of items having a DIF indicating that they are easier for Danish students at level B (darker yellow), level ICILS (medium yellow), and level C (lighter yellow). Orange bars show the number of items being harder for Danish students at levels B, ICILS, and C

for the remaining countries. The conclusion from these comparisons supported the hypothesis that the DIF could be considered uniform (Hanson 1998).

The last three columns in Table 1 report the number of items that showed DIF according to the ICILS criterion (DIF larger than half of 0.6 logits) and the MH D-DIF Level B (DIF larger than half of 0.638 logits) and Level C (DIF larger than half of 0.426 logits) criteria.

The number of items showing DIF is rather high, but this observation is not of primary interest for this study. In order to get insight into the content of the DIF items, the items

Table 3 Easier and harder items for Norwegian students

Easier	Harder
Aspect 1.1: Knowing about and understanding computer use	*Aspect 1.1: Knowing about and understanding computer use*
Open a link (− 0.23)	Save a presentation with a new file name (0.41)
Navigate to a text-based URL (− 0.22)	Open a link to a different page of a website (0.35)
Open a file of a specified file type (− 0.44)	*Aspect 1.2: Accessing and evaluating information*
Aspect 1.2: Accessing and evaluating information	Find specific information on a website (0.3)
Evaluate the reliability of a crowd sourced information website (− 0.46)	*Aspect 2.2: Creating information*
	Create a balanced layout for text and images in a website page (0.25)
Aspect 1.3: Managing information	
Modify the sharing settings of a collaborative document (− 0.34)	*Aspect 2.3: Sharing information*
	Range of relevant information on a topic included in a poster (0.31)
Aspect 2.1: Transforming information	
Adapt information for an audience (− 0.31)	*Aspect 2.4: Using information securely and safely*
Aspect 2.3: Sharing information	Differentiate paid search results from organic search results (0.31)
Identify who received an email by carbon copy (− 0.38)	
Aspect 2.4: Using information securely and safely	
Explain a potential problem if a personal email address is publicly available (− 0.24)	

Numbers in parenthesis are the DIF values

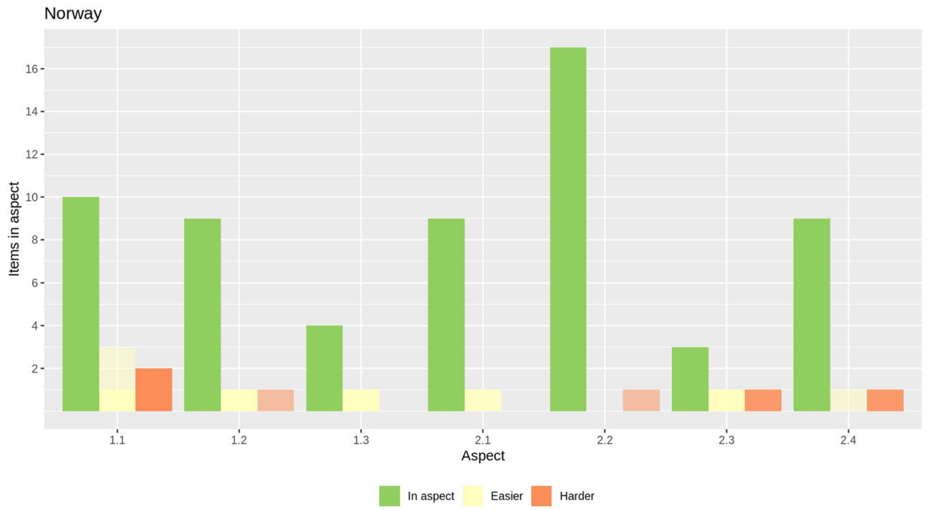

Fig. 2 Items showing DIF for Norwegian students. Green bars show how many items there are in the aspect in total. Yellow bars show the number of items having a DIF indicating that they are easier for Norwegian students at level B (darker yellow), level ICILS (medium yellow), and level C (lighter yellow). Orange bars show the number of items being harder for Norwegian students at levels B, ICILS, and C

were collected in groups based on the ICILS study's identification of the items in relation to the strands and aspects. As the Danish National Research Coordinator of ICILS.

I had access to a mapping of items onto aspects in the ICILS working documents (IEA Data Processing Center, IEA Secretariat, and The Australian Council for Educational Research 2014).

In Tables 2, 3 and 4, the description of the items is given together with the sizes of the DIF (Figs. 1, 2 and 3 show the sizes of DIF visually).

Table 4 Easier and harder items for German students

Easier	Harder
Aspect 1.1: Knowing about and understanding computer use Click on a link (— 0.28) Save a presentation with a new file name (— 0.39) Switch applications to an internet browser from the taskbar (— 0.27) *Aspect 1.2: Accessing and evaluating information* Find specific information on a website (— 0.22) *Aspect 2.2: Creating information* Design and layout text in a poster (— 0.3) Uses colour consistently throughout poster to convey meaning (— 0.27) Establish a clear and complimentary role for a photo and its description on a website page (— 0.32) *Aspect 2.4: Using information securely and safely* Explain the features that make one of two passwords more secure (— 0.28) Recognise that usage restrictions for images are a legal issue (— 0.22)	*Aspect 1.1: Knowing about and understanding computer use* Open a link to a different page of a website (0.53) *Aspect 2.1: Transforming information* 2.1: Excludes irrelevant information in a poster (0.35) *Aspect 2.4: Using information securely and safely* Identify information that is risky to include on a public profile (0.49) Identify that an email does not originate from the purported sender (0.64) Identify that a link's URL does not match the URL displayed in the link text (0.25)

Numbers in parenthesis are the DIF values

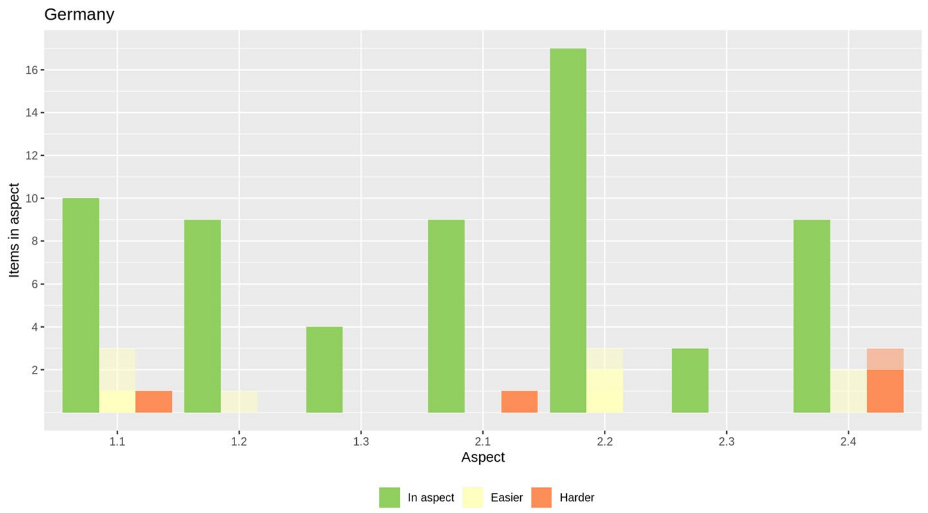

Fig. 3 Items showing DIF for German students. Green bars show how many items there are in the aspect in total. Yellow bars show the number of items having a DIF indicating that they are easier for German students at level B (darker yellow), level ICILS (medium yellow), and level C (lighter yellow). Orange bars show the number of items being harder for German students at levels B, ICILS, and C

Both Danish, German and Norwegian students find a number of items from Aspect 1.1, *Knowing about and understanding computer use*, easier than their peers of the same ability level from the other participating countries. Items in this aspect concern opening a link, navigating to URL's by inserting them into the browser address bar, opening files of specific types, and switching applications from the task bar. As can be seen from the descriptions, these items are connected to basic use of computers, and therefore address the computer literacy aspect of Computer and Information Literacy measured in ICILS.

The second observation is that Danish students find items from Aspect 2.1, *Transforming information*, easier than their peers from other countries. Some of the items from this aspect are related to computer literacy, like using software to crop an image, but most of them are more related to information literacy, like excluding irrelevant information in a poster, adapting information to an audience, and converting a description of directions into a visual route on a map. Norwegian students find a single item (adapt information for an audience) from Aspect 2.1 easier than peers of similar abillity in other participating countries.

German students, on the other hand, find one item (exclude irrelevant information in a poster) in Aspect 2.1 more difficult than their peers.

Thirdly, German students find two items from from Aspect 2.4, *Using information securely and safely*, easier, and three items harder than peers from other participating countries. The easy items are connected to information literacy (they test if students can identify features that make one of two passwords more secure, and recognize that usage restrictions for images are a legal issue), while two of the harder items are connected to computer literacy (identify that an email does not originate from the purported sender, and that a link's URL does not match the URL displayed in the link text). The third of the harder items are more connected to information literacy (identify information that it is risky to include in a public profile).

Danish students find an item from Aspect 2.4 easier than their peers in other countries. But, on the other hand, they find two items from this aspect more difficult than their peers. The easy item is connected to being able to understand technical aspects of secure Internet use (identify that an email does not originate from the purported sender). One of the more difficult items is of the same kind, namely the one that tests students' ability to identify URL fraud, while the other is about identifying information that is risky to include in a public profile, and could be said to be more related to information literacy.

Norwegian students also find an item in aspect 2.4 easier than their peers. This item, explaining the potential problem if a personal email address is publicly available, is more connected to information literacy. And the Norwegian students also find one item more difficult, namely the one related to identifying paid search results from among organic search results. This is considered more related to information literacy.

Two items in Aspect 2.2, *Creating information*, are harder for Danish students. These items are related to information literacy, more specifically to the layout of a presentation or information sheet, including laying out images and creating a balanced layout of text and images. Norwegian students also find the latter item harder.

Contrary to this, German students find three items from Aspect 2.2 easier. The items German students find easy in Aspect 2.2 are all related to layout, including designing and laying out text, using colors consistently, and establishing a clear role for a photo and caption on a website.

The final observation is related to Aspect 1.2, *Accessing and evaluating information*. Danish students find items from this aspect harder than their peers from other countries. The Danish students have trouble selecting relevant images in a presentation and presenting accurate information. These items are related to information literacy. The same goes for the item that the Norwegian students find difficult, namely finding information on a website. Contrary to this, German students find it easier to find specific information on a website.

Discussion

This paper shows that essential results can be identified by comparing the distribution of difficulties of items in international large-scale assessments. This is a more constructive approach to the challenge of DIF, but it does not eliminate the serious threat to the validity of the comparison of countries. One explanation for the DIF could be that the CIL construct is in fact more than one construct related to the two strands, collecting and managing information, and producing and exchanging information. This was partly the conclusion in the study by Punter et al. (2017) mentioned earlier, even though they split the items into three strands: evaluating and reflecting on information, communicating information, and applying technical functionality. This study underpins the hypothesis that CIL may be two things: Computer Literacy and Information Literacy. Therefore, I propose in future studies to investigate the psychometric properties of a two-dimensional scale composed of these two aspects.

While I believe that the content-oriented approach to DIF used in this paper provides very important knowledge, which could be used in large-scale international assessment studies to inform educators more about the content aspects of the assessment, I do also want to bring up some concerns.

First, even though I think I have identified a number of important insights, the DIF does show a somewhat unclear picture. One example is the items measuring computer literacy in Aspect 2.4 that Danish students found easier and harder, respectively. Second, the number of items showing DIF is rather small [even though it can be considered high when taking into account the severity of the DIF in relation to the league tables (cf. Kreiner and Christensen 2014)].

Conclusions

A number of conclusions can be drawn based on these observations. First, it seems that Norwegian and Danish students find items related to computer literacy easier than their peers from other countries. This could be connected to the fact that Denmark and Norway have some of the highest ICT Development Indexes worldwide, and that computers are highly available in their classrooms (Fraillon et al. 2014, p. 96). Students in these countries are used to working with computers, probably more than their peers from the other participating countries.

Second, however, Danish, and to a certain degree Norwegian, students find items related to information literacy more difficult. This is the case when it comes to the layout and design of posters, information sheets, etc., and when it comes to communicating appropriately in a specific situation.

Opposed to this, German students seem to be comparably good at designing and laying out posters, web pages etc.

From a Danish perspective, these results are rather surprising and alarming. Information literacy has been an integral part of the teaching and learning standards, especially in relation to teaching and learning Danish, for several years (Undervisningsministeriet 2009, 2014), and the use of computers for research has been promoted for decades (Bundsgaard et al. 2014, p. 111f.). If Danish students are struggling with assessing and evaluating, managing, and creating information, they will face problems in their future studies, as citizens, and at the workplace.

Do these conclusions indicate that having easy access to technology might help develop basic computer skills, while more critical parts of computer and information literacy need more focus in teaching practices to be developed?

As the title of this paper suggests, identifying country DIF in international comparative educational studies can be considered a pedagogical tool. The analyses can give teachers and curriculum designers knowledge of which aspects of a construct students in a specific country find particularly easy or hard, and this can be used in giving these particular aspects extra focus in teaching. Based on the analyses in this paper, a recommendation for Danish (and to a certain degree Norwegian) teachers would be to put extra emphasis on teaching information literacy, while German students might gain if their teachers put more emphasis on computer literacy.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40536-019-0077-2.

Additional file 1. R script used in the analyses.

Abbreviations

CIL: computer and information literacy; DIF: differential item functioning; ETS: *educational testing service*; IEA: *International Association for the Evaluation of Educational Achievement*; ICILS: *international computer and information literacy study*; MH D-DIF: Mantel and Haenszel's differential item functioning; PIRLS: *progress in international reading literacy study*; PISA: *Programme for International Student Assessment*; TAM: *test analysis modules*; TIMSS: *trends in international mathematics and science study*.

Acknowledgements

The author wishes to thank the editor Peter Van Rijn and the anonymous reviewers for the invaluable comments and suggestions to an earlier version of the paper. The author also would like to express his deep gratitude to the thousands of students and teachers around the globe who gave their time and effort by taking the test and answering questions.

Authors' contributions

The author read and approved the final manuscript.

Funding

This research was not funded.

Availability of data and materials

The data set used is freely available at https://www.iea.nl/repository/studies/icils-2013 after login (user creation is free). The R script used in the analyses is available as Additional file 1.

Competing interests

The author was National Research Coordinator for Denmark in ICILS 2013.

Received: 22 December 2018 Accepted: 21 August 2019
Published online: 11 September 2019

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/BF02293814.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum.

Bundsgaard, J., Pettersson, M., & Puck, M. R. (2014). *Digitale kompetencer. It i danske skoler i et internationalt perspektiv [Digital Competences. IT in Danish Schools in an International Perspective]*. Aarhus: Aarhus Universitetsforlag.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age. The IEA international computer and information literacy study international report*. Cham: Springer.

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International computer and information literacy study: Assessment framework*. Retrieved from http://ifs-dortmund.de/assets/files/icils2013/ICILS_2013_Framework.pdf.

Fraillon, J., Schulz, W., Friedman, T., Ainley, J., Gebhardt, E., Ainley, J., et al. (2015). International association for the evaluation of educational achievement (IEA). *ICILS 2013: Technical report*.

Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education, 30*(3), 178–195. https://doi.org/10.1080/08957347.2017.1316276.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes*. https://doi.org/10.1186/s12955-017-0755-0.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*(3), 244–253.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology, 36*(2), 378–390. https://doi.org/10.1080/01443410.2014.946890.

IEA Data Processing Center, IEA Secretariat, & The Australian Council for Educational Research. (2014, February). *International computer and information literacy study. Item map for CIL framework.*

Innabi, H., & Dodeen, H. (2006). Content analysis of gender-related differential item functioning TIMSS items in mathematics in Jordan. *School Science and Mathematics, 106*(8), 328–337. https://doi.org/10.1111/j.1949-8594.2006.tb17753.x.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9*(2), 122–133. https://doi.org/10.1080/15305050902880769.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale: Lawrence Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Mislevy, R. J. (1998). Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education, 11*(1), 49–63. https://doi.org/10.1207/s15324818ame1101_3.

OECD. (2014). *PISA 2012—technical report*. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing, 13*(3), 272–293. https://doi.org/10.1080/15305058.2012.738266.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023–1046. https://doi.org/10.1177/0013164411400734.

Punter, R. A., Meelissen, M. R. M., & Glas, C. A. W. (2017). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal, 16*(6), 762–780. https://doi.org/10.1177/1474904116672468.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks pædagogiske Institut.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules. R package version 2.8-21*. Retrieved from https://CRAN.R-project.org/package=TAM.

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing, 13*(2), 152–174. https://doi.org/10.1080/15305058.2012.690140.

Undervisningsministeriet. (2009). *Fælles mål 2009—Dansk [Common Goals 2009—Danish]*. Retrieved from http://www.uvm.dk/Service/Publikationer/Publikationer/Folkeskolen/2009/~/media/Publikationer/2009/Folke/Faelles%20Maal/Filer/Faghaefter/120326%20Faelles%20maal%202009%20dansk%2025.ashx.

Undervisningsministeriet. (2014). Forenklede Fælles Mål Dansk *[Simplified Common Goals 2014 Danish]*. Retrieved October 30, 2015, from EMU Danmarks læringsportal website: http://www.emu.dk/omraade/gsk-lærer/ffm/dansk.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ConQuest: Generalised item response modelling software (version 2.0)*. Camberwell: ACER Press.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*(3), 287–300. https://doi.org/10.1207/s15327574ijt0603_5.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika, 82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8.

Publisher's Note