Moving beyond the study of gender differences: An analysis of measurement invariance and differential item functioning of an ICT literacy scale

Ove E. Hatlevik, Ronny Scherer and Knut-Andreas Christophersen

University of Oslo, Norway

Author Note

Ove E. Hatlevik, Faculty of Education and International Studies, Oslo and Akershus University College of Applied Sciences; Ronny Scherer, Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo; Knut-Andreas Christophersen, Department of Political Science, Faculty of Social Sciences, University of Oslo.

Correspondence concerning this article should be addressed to Ove E. Hatlevik, Faculty of Education and International Studies, Oslo and Akershus University College of Applied Sciences, PO box 4 St. Olavs plass, 0130 Oslo, Norway, E-Mail: ove-edvard.hatlevik@hioa.no.

Abstract

Crafting a validity argument is crucial for the development of any assessment of ICT literacy. In about the context of studying gender differences in ICT literacy, it has therefore become essential to ensure that gender differences are not due to the existence of measurement bias, which might indicate that an assessment instrument used to measure ICT literacy operates differently for girls and boys. Hence, researchers need to gather evidence on the validity of such gender comparisons. The present study follows this line of research by investigating the overall measurement invariance at the construct level and the differential functioning of items across gender of an ICT literacy test. Based on the data obtained from a random sample of 919 Norwegian lower secondary school students (468 girls), multi-group confirmatory factor analysis showed that the test was invariant to a sufficient degree, and girls outperformed boys in the overall test score ($\beta = .35$, $p < .001$). Yet, differential item functioning existed for selected items. These results highlight the importance of testing for measurement invariance and differential item functioning that goes beyond the mere description of gender differences. Moreover, attention is brought back to the validity of ICT literacy assessments, and ways to improve these assessments are discussed.

Keywords: Differential item functioning; Gender; ICT literacy; Lower secondary students; Measurement invariance

Moving beyond the study of gender differences: An analysis of measurement invariance and differential item functioning of an ICT literacy scale

1. Introduction

The concept of information and communication technology (ICT) literacy describes what students can do and master with digital technology such as computers, tablets, and smart phones. Inspired by the 21st century framework (Binkley et al., 2011), ICT literacy can be defined as the ability or capacity "*to solve problems of information, communication and knowledge in digital environments*" (Claro, 2012, p. 1043). Aesaert et al. (2015) argued that this definition provides a broader understanding of ICT literacy than earlier definitions of ICT literacy emphasizing digital skills and knowledge about how the computer works. Initially, the term literacy referred to the "ability to write and read written language" (Kin, Kil, & Shi, 2014, p. 29). Kin and colleagues (2014) argue that, due to the emergence of a digitalized language, the meaning of literacy was broadened to cover the ability to read and write language in a digital context (Kin et al., 2014).

Because of the focus on ICT literacy, there has been an increase in the number of publications on tests measuring what students can achieve when they use ICT in the past decade (Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016). These tests were developed and used in different countries, such as Australia (Ainley, Fraillon, & Freeman, 2007), Chile (Claro et al., 2012), Italy (Calvani, Fini, Ranieri, & Picci, 2012), Korea (Kim, Kil, & Shin, 2014), Norway (Hatlevik, Ottestad, & Throndsen, 2015), and the United States of America (Educational testing service, 2002; Huggins, Ritzhaupt, & Dawson, 2014). Clearly, these tests differed with respect to their test content, language, psychometric quality, and the ways in which a validity argument has been created (Siddiq et al., 2016). For instance, although the generalizability of the test scores across sub-groups of students is an important argument for the construct validity of any

measure (AERA, APA, & NCME, 2014), only a limited number of studies reported evidence to support such an argument for measures of ICT literacy. Among the probably most controversially discussed sub-group differences are gender differences (Hohlfeld, Ritzhaupt, & Barron, 2013; Volman, van Eck, Heemskerk, & Kuiper, 2005).

Gender differences are often examined in the context of students' use and experience of ICT (OECD, 2015; Volman et al., 2005), their attitudes toward ICT (Imhof, Vollmeyer, & Beierlein, 2007; Whitley Jr., 1997), computer self-efficacy (Lau & Yuen, 2015; Tømte & Hatlevik, 2011), and their performance on ICT-related tasks (Aesart & van Braak, 2015). However, one major assumption of examining gender differences is often not explicitly addressed, that is, the extent to which the test scores obtained from the ICT-related measures are comparable across gender. In other words, it often remains unclear whether the measures work equally well for girls and boys. Yet, neglecting this assumption of what is called "measurement invariance" is troublesome, because major violations of it and therefore large variations in how a measure works may lead to considerable bias in gender comparisons (Millsap, 2011). This measurement bias may have caused the inconsistent findings on gender differences across studies[1]: whereas some studies identified gender differences in favour of boys (Rubio, Romero-Zaliz, Mañoso, & de Madrid, 2015), others showed that girls outperform boys (Padilla-Meléndez, del Aguila-Obra, & Garrido-Moreno, 2013); moreover, in others no significant difference was found (Silva-Maceda, Arjona-Villicaña, & Castillo-Barrera, 2016). These inconsistencies warrant a deep analysis of the methods used to examine the gender gap.

Addressing this concern, the present study examines the invariance of an ICT literacy measure across gender, and investigates specific variations of how this measure

[1] We would like to thank one anonymous reviewer for his or her input on the inconsistencies of findings on this topic.

works across the two groups (i.e., differential item functioning). If a sufficient degree of invariance can be established, gender comparisons of the resultant ICT literacy scores will be conducted.

1.1 The Concept of ICT literacy

ICT literacy is one of many concepts used to describe students' capabilities to use and perform with digital technology. These concepts often include descriptions of capabilities (i.e., skills, competences, or literacy) within the context of digital technology (i.e., information, computer, ICT or digital). Ferrari (2013) reviewed existing frameworks of digital competence and ICT literacy and found that the use of terms can be traced back to different research traditions and countries. For instance, it seems that the concept of ICT literacy has been more used in Asia and the US, whereas the concept of digital competence has been more used in Europe. Albeit these differences in the use of a specific terminology seems to be almost arbitrary, differences in the dimensions of what is called ICT literacy or digital competence exist (Siddiq et al., 2016).

In 2002, the Educational Testing Service (ETS) defined ICT literacy as students' ability to "access, manage, integrate, evaluate, and create information to function in a knowledge society" (p. 2). Building on this definition, the OECD initiated a feasibility study where Lennon, Kirsch, Von Davier, Wagner and Yamamoto (2003, p. 8) referred to ICT literacy as "the interest, attitude and ability of individuals to appropriately use digital technology and communication tools" (see also Martin & Grudziecki, 2006, p. 251). Both definitions include five similar areas of ICT literacy: the ability to (a) access information, (b) manage information, (c) integrate information, (d) evaluate information, and (e) create information. In this way, the concept of ICT literacy covers a wide range of competences. Even further, building on these descriptions (Educational Testing Service, 2002; Schleicher, 2008), other aspects have been added to the concept

of ICT literacy, such as digital problem solving, collaboration, technical operations, ethics, and responsibility (Binkley et al, 2012; Ferrari, 2012). Ferrari (2012, 2013) observed that many definitions further substantiate a specific goal that is associated with the acquisition of ICT literacy or digital competence, that is, "to participate effectively in society" (Lennon et al., 2003, p. 8; Schleicher, 2008, p. 632), "to function in a knowledge society" (Educational Testing Service, 2002, p. 2), and to use digital technology "efficiently and responsibly" (Norwegian Directorate for Education and Training, 2012).

Taken together, these considerations point to (a) the importance of ICT literacy for an active and responsible participation in our information society; (b) the variety of definitions of the concept, particularly with respect to their specific focus; (c) the fact that ICT literacy comprises several competences students should acquire as active citizens. In the current study, we broadly refer ICT literacy to the capacity to solve a variety of problems with different difficulty in a digital context (Claro et al., 2012). This understanding of ICT literacy ranges from mastering ICT applications to higher-order thinking in a digital environment, and includes preparing students for continuous learning (Claro et al., 2012, p. 1043). Moreover, our conceptualization of ICT literacy is also informed by a national framework "*developed to serve as a reference document for developing and revising the National Subject-Specific Curricula*" (Norwegian Directorate for Education and Training, 2012, p. 5), which differentiates between four sub-categories of the concept ICT literacy: search and process, produce, communicate, and digital judgement (Norwegian Directorate for Education and Training, 2012).

1.2 Gender differences in ICT literacy

In the context of gender differences, it is important to distinguish between self-reported ICT literacy and performance-based ICT literacy assessments. When students

rate their own ICT literacy, it seems that boys are reporting higher levels of ICT literacy compared to girls (Lau & Yuen, 2015; Litt, 2013). One limitation with self-reports is that the ratings are influenced by how they perceive themselves and their capabilities (Rohatgi, Scherer, & Hatlevik, 2016).

Despite the existence of many studies on gender differences in ICT literacy, as measured by performance tests, the existing body of literature abounds in inconsistent findings. In some studies, boys outperform girls (Calvani et al., 2012; Gui & Argentin, 2011; Van Deursen, 2012), whereas in other studies, the opposite results appeared (Fraillon et al., 2014; Claro et al., 2012; Kim et al., 2014; Kim & Lee, 2011; Yang, 2012). In addition, some studies could not identify any significant gender differences in ICT literacy (Hargittai & Shafer, 2006; Hatlevik & Christophersen, 2013; Van Deursen, Van Dijk & Peters, 2011; Van Deursen & Van Dijk, 2009). These findings give rise to the question about potential reasons that may explain the substantial variation in gender effects. From a sociological perspective, gender differences in performed or perceived ICT literacy may be due to students' perceptions of gender roles, their attributions to failure, and the extent to which stereotypical expectations (e.g., "Computer science and mathematics are boys' domains") prime their performance and attitude toward tasks that require ICT (i.e., "stereotype threat"; Koch, Müller, & Sieverding, 2008; Sieverding & Koch, 2009). From a methodological perspective, group comparisons rely on the assumption that test items and the entire ICT literacy scale operate equally across gender (Millsap, 2011). Potential deviations from what is called measurement invariance, for instance, in a sense that items may function differently (Holland & Wainer, 1993), may cause bias in the estimated scale scores across groups and therefore threaten the validity of gender comparisons. To rule out that variations in the

functioning of an ICT literacy test and the corresponding items exist across gender, testing for measurement invariance and differential item functioning is necessary.

1.3 Measurement invariance (MI) and differential item functioning (DIF)

When developing an instrument to assess ICT literacy, obtaining evidence on how the instrument is operating in a sample or sub-groups is critical – if group comparisons are to be valid, it needs to be ensured that the test instrument works in the same way across groups, and that the ICT literacy construct has the same theoretical structure (Dimitrov, 2010). In this section, we will outline two ways of approaching comparability across groups, which have been largely applied in educational research: MI testing based on a multi-group modelling approaches and testing for item-specific DIF. These approaches are powerful tools in the methodological toolbox, as they enable researchers in the field or educational technology to address critical aspects of validity (Teo, 2015; Teo, Lee, Chai, & Wong, 2009).

Multi-group invariance testing approaches. Testing for measurement invariance with the help of multi-group approaches is based on a series of models that differ in the extent to which specific parameters in the measurement model of ICT literacy (e.g., factor loadings, item intercepts, item residual variances) are constrained to equality across groups of students. Brown (2006) distinguishes between different types invariance models in the multi-group approach: The starting point for a MI analyses is to examine whether the structure of the theoretical factor model is supported in all groups (configural invariance). This invariance model assumes that the number of factors and the pattern of factor loadings are the same across groups; yet, no further similarities in model parameters are enforced. If this model holds for both groups, further constraints to the multi-group model can be added. In the second step of invariance testing, factor loadings are constrained to be equal across groups (metric invariance). This model

places the metric of the latent variables on the same scale for girls and boys. However, Schroeders and Wilhelm (2011) argued that constraining only factor loadings to equality across groups is not sufficient to conduct meaningful comparisons of factor means across groups when response data are categorical. Hence, they suggest moving beyond metric invariance by further constraining the item thresholds/intercepts (scalar invariance; see also Brown, 2006). If this model fits the data reasonably well, latent variables have the same meaning across groups, and "potential differences in the means of the latent variables are interpretable" (Schroeders & Wilhelm, 2011, p. 901). Finally, the third step adds constraints to the residual variances of items (strict invariance). This represents the most restrictive model; if it holds and a reasonable fit to the data is achieved, the model ensures that the construct is measured with the same reliability across groups (Schroeders & Wilhelm, 2011). Yet, although at least scalar invariance should be established to make group comparisons meaningful (Millsap, 2011), Demitrov (2010) pointed out that there is neither perfect invariance nor evidence of complete inequality. Consequently, some restrictions of model parameters on specific items may be relaxed without losing the comparability of measures (van de Schoot et al., 2013). Notice that items, for which constraints can be relaxed, represent potential threats to measurement invariance, as their corresponding parameters in the measurement model are not equal across groups. Such deviations from invariance that exist for specific items are broadly referred to as differential item functioning (DIF; Millsap, 2011). In this context, we note that the proposed invariance testing procedure by means of multi-group modelling is conducted for the overall measurement model, whereas item DIF testing focuses on the deviations from measurement invariance for specific items. Researchers can thus obtain information on the extent to which, for instance, a hypothesized factor structure applies to groups within the sample, thereby interrogating

the data for evidence on construct validity (Pellegrino, DiBello, & Goldman, 2016). However, multi-group modelling approaches do also allow for testing the extent to which specific items function differently across groups, yet, with a considerably lower sensitivity than direct approaches, such as the MIMIC-DIF approach, which will be outlined subsequently (Bauer, 2016).

Testing for item-specific differential item functioning (DIF). De Ayala (2009) stated that items that exhibit DIF work "one way for one group of respondents and in a different way [for another group]" (Ayala, 2009, p. 323). In other words, "two individuals of similar ability do not have the same probability of answering a question [or item] in a particular way" (APA, 2014, p. 93). Students' performance on an item, conditional on their ability or latent trait, is therefore dependent on the group they belong to. This deviation from measurement invariance can have different forms. For instance, so-called "uniform DIF" describes situations in which the probability to answer an item correctly is consistent across all levels of abilities for one subgroup. Such an item works differently for various groups, and this difference is equal at all levels of ability. In contrast, "non-uniform DIF" occurs if an item works differently for various groups, but the difference depends on the level of the latent trait. Hence, there is an interaction between group membership and the latent variable or ability. Put differently, the probability to succeed on an item could be lower or higher for a subgroup for some ability levels and not for other ability levels.

Identifying potential DIF means to "examine the instrument at the item level to see whether one or more items may be considered biased" (De Ayala, 2009, p. 324). There are several strategies to detect DIF, and the Multiple-Indicators-Multiple-Causes (MIMIC) approach represents one of them (Bauer, 2016; Woods, 2009). MIMIC models represent latent variable models that are comprised of a measurement model and a

structural part that connects a covariate (e.g., gender, age) to a latent variable and a specific item (Brown, 2006). To uncover potential *uniform DIF* for a single item using a MIMIC model, at least two steps need to be taken. First, a baseline model is specified which comprises the measurement model and the covariate. A direct effect of the covariate on the latent variable is introduced, and direct effects of the covariate to the items are constrained to zero (Chun, Stark, Kim, & Chernyshenko, 2016). Second, a series of MIMIC-DIF models are specified, in which one item is regressed on the covariate at a time. These direct effects reflect on potential differences in the item difficulties (thresholds) across the values of the covariate, yet not effects on item discriminations (factor loadings). Figure 1a depicts the general case of these models. These models are often referred to as "augmented models", and are compared to the baseline model with respect to their model fit (Woods, Ottmann, & Turkheimer, 2008). An item is flagged with uniform DIF if the augmented model fits the data significantly better than the baseline model ($p < .05$) and the direct item effects are statistically different from zero (Chun et al., 2016). Chi-square difference or likelihood-ratio tests are used to conduct these model comparisons. At this point, it must be noted that alternative approaches have been proposed to detect uniform DIF with MIMIC models; these include approaches that are based on modification indices or different baseline models (Woods, 2009). Following Chun et al.'s (2016) recommendations, the approach taken in this study seemed most efficient in terms of its power to detect uniform DIF items.

The MIMIC-DIF testing approach can also be used to detect *non-uniform DIF*. Building on the augmented models to detect uniform DIF, the interaction between the latent variable and the covariate is added to these models (see Figure 1b). An item is flagged with non-uniform DIF if the resultant models fit the data significantly better than

the uniform DIF models and the interaction effect is statistically different from zero (Chun et al., 2016).

There are some similarities between DIF testing and the multi-group measurement invariance testing approach because they both deal with bias or fairness in testing. However, they work in different ways, as DIF focuses on the item level, whereas the multi-group approach focuses on the overall test level. Both procedures may therefore complement each other (Bauer, 2016).

1.4 Aims of the present study and research questions

This study aims to assess the overall measurement invariance and potential differential item functioning of an ICT literacy measure across gender. Moreover, if a sufficient degree of invariance can be achieved (i.e., scalar invariance), mean differences in students' ICT literacy can be examined. Specifically, the following research questions are addressed:

1. To what extent does the measure of students' ICT literacy show overall invariance across gender? (*Multi-group modelling*)
2. To what extent do uniform and non-uniform differential item functioning across gender exist for the items measuring students' ICT literacy? (*Differential item functioning*)
3. If a sufficient level of invariance can be established, how do girls and boys differ in their performance on the ICT literacy test? (*Mean comparisons*)

2. Methods

2.1 Participants and procedure

The selection of students for this cross-sectional study followed two adjacent steps: First, 145 schools with 9th grade students, who were between 14 and 15 years old,

were randomly selected from a national list of schools. Geographical location, school size, and school type were used as strata. In the second step, school principals were asked to randomly choose one class of students to participate in a web-based test of ICT literacy. Principals then received an email with information regarding how to give students access to the web-based test. The schools were asked to set aside time for students completing the test during lesson hours at school. A total of N = 919 students from 53 schools participated in the study (age: 14-15 years; 50.9% female), yielding a total response rate of approximately 37% at the school level. The Norwegian Data Protection Authority (www.datatilsynet.no) was notified in advance about the study, and the data collection was carried out based on their guidelines. All participating schools volunteered and gave their consent to use the data for scientific purposes; students' identities were completely anonymized.

2.2 Instruments

The students were asked to complete a multiple-choice test consisting of 14 questions (see Table 1) that were developed based on the Norwegian ICT literacy curriculum and related to themes such as searching and processing, producing, communicating, and evaluating digital information (Norwegian Directorate for Education and Training, 2012). The topics mentioned in the national framework – as they were operationalized as competence goals – were expected to be covered through classes at school. The resultant test focused more on ICT-related knowledge as a facet of ICT literacy rather than the performance of specific skills or competences. Each of the 14 multiple-choice questions had four response options, one of which was correct. Items were scored automatically (i.e., by computer-based means) as either correct (code: 1) or incorrect (code: 0). Non-responses were coded as missing. All items were administered

in Norwegian. Please find the item stimuli and response options in the Supplementary Material S4.

The development of the ICT literacy was informed, on the one hand, by the existing frameworks of ICT literacy – or what is often called "digital competence" (Siddiq et al., 2016) – and, on the other hand, the development of an ICT school curriculum in Norway. These two sources provide information about the processes or core dimensions of ICT literacy and possible progressions students might show across schooling and along a curriculum. Moreover, they inform teachers and principals about the ways to foster students' ICT literacy as a transversal skill (Greiff et al., 2014).

Since 2006, the Norwegian school curricula have defined subject- and grade-specific competence goals, in which the concept of ICT literacy is integrated in the existing subjects – it is thus understood as a transversal skill. These competence goals formulate specific expectations about the use of ICT, such as "Using word processing tools for archiving and cataloguing their [students'] own work". Later in 2012, the existing competence goals were refined and resulted in a framework for digital skills that served as a "reference for developing subject and grade relevant competence aims" (Norwegian Directorate for Education and Training, 2012, p. 5). As mentioned earlier (of section 1.1), this framework proposed the following dimensions (sub-categories) of digital skills: (1) Search and process, (2) Produce, (3) Communicate, and (4) Digital judgement. 'Search and process' includes searching for information, navigating information searches, evaluating information, as well as sorting out, categorizing, and interpreting information. 'Produce' refers to composing, reapplying, converting, and developing digital elements, as well as adhering to copyright. 'Communicate' refers to the use of ICT for collaborating with peers in the classroom, as well as presenting knowledge and information to specific audiences. 'Digital judgement' covers the

responsible and ethical use of ICT, including considering privacy and regulations. The following list provides examples of items used in the ICT literacy test that are mapped onto this framework:

- Items Q1 and Q2 are examples of the sub-category 'Search and process': Q1 tests knowledge about digital tools to register data from fieldwork. Q2 measures knowledge about using both online and digital maps.
- Items Q12 and Q14 are examples of the sub-category 'Produce': Q12 tests knowledge about copyright rules for using sources. Q14 measures knowledge about the necessity to provide references to information sources.
- Items Q4 and Q11 are example of the sub-category 'Communicate': Q4 tests knowledge about how students can contribute and develop an opinion with digital tools. Q11 measures the skill to identify the appropriate tool for communication.
- Items Q6 and Q9 are examples of the sub-category 'Digital judgement': Q6 tests knowledge about the publication of incorrect information online. Q9 measures knowledge about privacy rules.

2.3 Analytical strategy

Before addressing our research questions, the data were first analysed with respect to their descriptive statistics (% correct item responses). Second, scale reliability was examined using the Kuder-Richardson reliability coefficient (KR-20; Acock, 2012) for dichotomously scored item responses. Third, item factor loadings were examined to obtain information about the quality of the internal structure of the scale and therefore the appropriateness of the measurement model. It was desired that all factor loadings are significantly different from zero on a 5% level, and preferably exceed 0.20 or even higher values such as 0.40 (Crocker & Algina, 2006; McAlpine, 2002).

To examine the goodness-of-fit of the underlying measurement model, the following indices were used: the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), the Weighted Root Mean Square Residual (WRMR) (Brown, 2006; Kline, 2016). The chi-square statistic was also examined; yet, this statistic is sensitive to the sample size such that the chi-square values are more likely to be significant in large samples. CFI values equal to or above 0.90, RMSEA values below or equal to 0.08, and WRMR values close to 1 indicate a good model fit (Marsh, Hau, & Grayson, 2005). Yet, we notice that these criteria do not represent "golden rules" and therefore strict criteria; they operate more as guidelines (Marsh, Hau, & Wen, 2004). The overall properties of the measurement model and potential misspecifications need to be considered as well (Brown, 2006). Furthermore, the WRMR index is still regarded as an "experimental" fit index, which has not yet been studied in detail. Hence, the guidelines concerning this fit index cannot be applied plainly (Yu, 2002).

Third, to address our first research question on the measurement invariance of the ICT literacy scale, we conducted multi-group confirmatory factor analysis (MG-CFA) with cumulative restrictions in the model parameters across gender; we tested for configural, scalar, and strict invariance. All models were estimated in the statistical software package M*plus* 7.1 (Muthén & Muthén, 1998-2015) with the Weighted Least Squares Means and Variance adjusted (WLSMV) estimator. Since items have been scored dichotomously in the present study, we chose WLSMV estimation (Beauducel & Herzberg, 2006), but supplemented this procedure in the context of DIF testing by robust maximum likelihood estimation (MLR). Following Schroeders' and Wilhelm's (2011) recommendation, we used the theta parameterization to avoid interpretation problems with respect to factor loadings, factor variance, and residual variance. The resultant invariance models were compared with respect to their chi-square statistics,

CFI and RMSEA. If the change in CFI is below .01 and the change in RMSEA is below .01, then the scalar model is as good as the configural (Cheung & Rensvold, 2002; Meade et al., 2008). Missing values for items ranged between 0.8% (Item Q7) and 5.7% (Item Q10). In the WLSMV estimation, missing values were handled by the pairwise deletion method (Asparouhov & Muthén, 2010); in the MLR estimation, missing values were handled by the Full-Information-Maximum-Likelihood procedure (Enders, 2010).

Fourth, we approached research question 2 with the MIMIC-DIF testing approach. To examine the extent to which uniform item DIF existed, we followed Chun et al.'s (2016) recommendations and specified a baseline model first. This model only contained a direct effect of gender on the latent variable (i.e., ICT literacy). Second, for each ICT literacy item, a MIMIC-DIF model was estimated that contained a direct effect of gender on the item under investigation (see Figure 1a). Given that the existing body of methodological research does not exhibit clear-cut findings on the most preferable estimation procedure in the context of detecting DIF with MIMIC models (e.g., Meade et al., 2008), we decided to specify the MIMIC-DIF models with the WLSMV estimator first, and with the robust maximum likelihood estimator (MLR) second. In the case of the WLSMV estimation, the baseline and augmented models are compared using chi-square differences testing; in the case of the MLR estimation, likelihood-ratio tests are performed (Woods et al., 2008). The resultant DIF effects and model comparisons are then compared across estimation procedures. This approach specifically accounts for potential effects of the estimation procedure on the detection of uniform DIF with MIMIC models and can be regarded as a robustness check of our findings. In fact, we consider this robustness check to be important, because it puts to test potential bias that might be due to the choice of estimators (Duncan et al., 2014).

Non-uniform DIF was examined by adding an interaction term between the latent variable (i.e., ICT literacy) and gender to the uniform DIF models (see Figure 1b). Model comparisons are then performed between the resultant models and the uniform DIF models. For this approach, only the MLR estimation was performed, because interaction terms between a latent and a manifest variable are hardly identified in WLSMV estimation (Muthén & Muthén, 1998-2015). Please find the corresponding M*plus* sample codes in the Supplementary Material S1-S3.

Fifth, if at least scalar invariance can be established with only few items that show differential item functioning (i.e., partial scalar invariance); factor means can be compared (Elosua & Mujika, 2015; Research Question 3). Commonly, one group serves as the reference, whereas the means of the other group is estimated.

We notice that the analysis of measurement invariance using a multi-group CFA approach (RQ1) was supplemented by testing for item-specific DIF (RQ2) to not only describe global deviations from invariance and the extent to which a series of parameter constraints affect the unequally sensitive changes in goodness-of-fit indices, but to obtain specific information about which items function differently across gender. Moreover, MG-CFA models are generally less parsimonious and may camouflage deviations from invariance for specific items (Brown, 2006).

3. Results

3.1 Item difficulties and scale reliability

Item easiness parameters of the ICT literacy test are indicated by the percentage of students who successfully solved the items (see Table 1). These parameters ranged between 26% (i.e., 26% of the students answered this item correctly) and 91% (i.e., 91% of the students answered this item correctly), and thus cover a sufficiently large spectrum. Hence, whereas item Q6 (item easiness = 91%) was the easiest item, item Q14

(item easiness = 26%) was the most difficult. In this sense, the less successful students solved an item, the higher the item difficulty.

The scale reliability for the entire sample was acceptable, KR-20 reliability = .67. Differentiating between girls and boys revealed comparable reliabilities of .66 and .68. These values can be regarded as acceptable for a test with dichotomously scored items that covers a broad construct.

3.2 Measurement model

In the first step, we tested whether a unidimensional CFA model fitted the data of the entire sample. Under the WLSMV estimation, this model fitted the data well, χ^2 [77] = 122.56, $p < .001$, CFI = 0.960, RMSEA = 0.025, 90%-CI RMSEA = [0.017, 0.034], WRMR = 0.972. All 14 items showed statistically significant factor loadings between 0.25 and 0.69 ($p < .01$; Table 1). This model formed the baseline for investigating measurement invariance under the multi-group approach and was further extended for uniform MIMIC-DIF models.

Under the MLR estimation, the factor loadings showed a similar range, λ = .24-.68; moreover, the information obtained from the test and its resultant scores was broadly spread and showed the maximal information between -2 and 0 logits (Figure 2), that is, for test scores around the mean performance, M = 8.94 (SD = 2.61). This model indicated a reasonable expected-a-posterior reliability of .65, and good item fit with Infit values between 0.99 and 1.00 and Outfit values between 0.99 and 1.04 (see De Ayala, 2009, for a detailed description of these item fit values). Moreover, the Q3 statistics – statistics that are based on residual correlations and indicate the extent to which item dependencies occur – ranged between -0.14 and 0.13 (M = -0.05, SD = 0.05), indicating that item dependencies existed only to a minor extent (Yen, 1984). The maximal chi-square value of the model was reasonable (χ^2 [91] = 13.17, p = .03), and additional fit

statistics supported the global fit of the baseline model, SRMR = 0.027, SRMSR = 0.036 (see Liu & Maydeu-Olivares, 2014, for a detailed description of these indices). This two-parameter logistic item response theory model was extended to uniform and non-uniform MIMIC-DIF models to approach RQ2.

Based on these psychometric properties of the ICT literacy test, we accepted a unidimensional model as the measurement model for the total student sample.

3.3 Multi-group confirmatory factor analyses (Research Question 1)

First, we tested for configural invariance, estimating all model parameters freely for boys and girls. This model resulted in a reasonable fit, χ^2 [154] = 230.05, $p < .01$, CFI = 0.932, RMSEA = 0.033, 90%-CI RMSEA = [0.024, 0.041], WRMR = 1.348. Given the reasonable fit of this model to the data, we further constrained the factor loadings and the item thresholds to achieve scalar invariance across gender in a second step. The resultant model exhibited to a reasonable model fit, χ^2 [166] = 241.81, $p < .01$, CFI = 0.932, RMSEA = 0.032, 90%-CI = [0.022, 0.040], WRMR = 1.424. Third, we constrained all factor loadings, thresholds, and residual variances to be equal across gender. This strict invariance model indicated a substantial loss in goodness-of-fit in comparison to both the configural and the scalar invariance model (see also Table 2), χ^2 [180] = 293.12, $p < .01$, CFI = 0.899, RMSEA = 0.037, 90%-CI = [0.029, 0.045], WRMR = 1.631. Considering the differences in goodness-of-fit statistics between adjacent invariance models (Table 2), we accepted the scalar invariance model, which showed an acceptable fit and only small changes in fit statistics after restricting the item thresholds. Hence, as a response to our second research question, we argue that scalar invariance across gender is defendable for the ICT literacy test.

3.4 Differential item functioning tests (Research Question 2)

Uniform DIF testing. As noted earlier, the detection of uniform DIF requires the specification of a baseline model. The baseline model in the case of WLSMV estimation showed a reasonable model fit, χ^2 [90] = 192.06, $p < .001$, CFI = 0.915, RMSEA = 0.035, 90%-CI [0.028, 0.042], WRMR = 1.165. Comparing the augmented models that contained a direct effect of gender on single items revealed that seven out of fourteen items functioned differently between girls and boys (Table 3). For these items, both the model comparisons to the baseline model and the uniform DIF effects were statistically significant. Yet, the directions of gender DIF varied across items; whereas items Q1, Q2, Q9, and Q13 were significantly easier for boys (β's < 0); items Q4, Q7, and Q11 were significantly easier for girls.

To test the robustness of these findings, we re-ran these analyses applying robust maximum likelihood estimation. Following the same procedure as under the WLSMV estimation, we added the direct gender effect on the latent variable (i.e., ICT literacy) to the measurement model (please refer to section 3.2). This model formed the baseline (*LL* = -6595.1, Npar = 29, SCF = 1.023, AIC = 13248, BIC = 13388, aBIC = 13296), to which the direct item effects were added. The models with item-specific, uniform DIF effects were then compared to the baseline model; the resultant fit statistics (i.e., information criteria) and model comparisons are shown in Table 4. The model comparisons and the DIF effects replicated the findings obtained from the WLSMV estimation procedure, both in terms of which items were flagged with uniform DIF and the directions of the gender differences. Hence, there is evidence on the robustness of the DIF analyses.

Non-uniform DIF testing. Finally, we tested for non-uniform DIF using MLR estimation. Overall, none of the fourteen items exhibited non-uniform DIF, as indicated

by insignificant differences in the log-likelihood values between the baseline model and the models specifying non-uniform DIF and the insignificant interaction effects of *ICT Literacy × Gender* on item responses (see Table 5). At this point, we would like to highlight that non-uniform DIF has been examined using an interaction variable between the latent variable ICT literacy and the manifest, categorical variable gender. Significant effects of this interaction variable on item intercepts (i.e., difficulties) indicate that the relation between gender and item intercepts is moderated by the level of ICT literacy; in other words, the extent to which gender DIF exists depends on the level of ICT literacy – the signs of the effects indicate the direction towards higher or lower levels. At the same time, significant interaction effects suggest that the relation between ICT literacy and item intercepts are moderated by gender (coded as *0 = boy*, *1 = girl*). Positive effects point to stronger relations for girls, whereas negative effects point to weaker relations for girls. Technically, the interaction variable *ICT literacy × Gender* was created using the XWITH command in M*plus* (Kline, 2016).

Taken together, the MIMIC-DIF testing procedure flagged seven items with uniform DIF across gender.

3.5 Gender differences in the level of ICT literacy (Research Question 3)

The third research question concerned gender differences in the level of ICT literacy among ninth graders. Given that scalar invariance was met, factor mean comparisons were meaningful. Yet, it must be noted that a considerable number of items showed uniform DIF; hence, mean comparisons should be based on a model that accounts for these DIF effects. We therefore specified a model that contained the direct effects of gender on the latent variable and the DIF effects. In the case of WLSMV estimation, for which only uniform DIF could be examined, this model fitted the data well, χ^2 [83] = 130.67, $p < .01$, CFI = 0.960, RMSEA = 0.025, 90%-CI [0.016, 0.033],

WRMR = 0.950. The unstandardized regression coefficient of gender on the latent variable was $B = 0.195$ ($SE = 0.057$, $p < .01$, $\beta = 0.354$ in the STDY standardization) and indicated that girls outperformed boys in the ICT literacy test. This finding was supported by the corresponding model that was based on MLR estimation (LL = -6565.5, Npar = 36, SCF = 1.0143, AIC = 13203, BIC = 13377, aBIC = 13262), in that we found the same direction of gender differences and effect size (β in the STDY standardization), $B = 0.371$, SE = 0.102, $p < .001$, $\beta = 0.365$. These findings converge to the existence of a "gender gap" in ICT literacy in favour of girls.

4. Discussion

There is an ongoing discussion about gender differences in students' ICT literacy (Fraillon et al., 2014; Lau & Yuen, 2015; Litt, 2013). To examine gender differences, it is necessary to consider that there is no bias in the tests used to measure ICT literacy. This can be done by investigating how the assessment instrument operates for girls and boys. The present study has taken this road and was aimed at examining the extent to which measurement invariance across gender held for an assessment of ICT literacy.

The first research question concerned the levels of measurement invariance that could be met across gender. The multi-group analyses showed that a scalar invariance model, with constrained model structures, factor loadings, and thresholds/intercepts across gender, was supported by the data. From a test developer's perspective, this finding strengthens the creation of a validity argument that addresses the generalizability of a factor structure across groups (AERA, APA, & NCME, 2014). In this sense, the current study provided evidence for the generalizability. From a practical point of view, this outcome is sufficient to conduct meaningful mean comparisons (Schroeders & Wilhelm, 2011). Based on the fit of the scalar model, the tested concept of ICT literacy seems to work equally for both boys and girls. However, the measurement

invariance testing procedure (i.e., MG-CFA) exhibiting scalar invariance focuses on the overall measurement model of ICT literacy, yet not necessarily the specific items that might deviate. This procedure should therefore be supplemented by differential item functioning analyses on the item level (Millsap, 2011).

The second research question dealt with the extent to which differential item functioning across gender existed for the items measuring students' ICT literacy. The DIF approach has a focus on the items in the ICT literacy test. Our DIF analyses show that eight items were subject to gender DIF. More specifically, it was more likely to succeed on five items for boys than for girls who had the same ICT literacy; the opposite direction was apparent for three out of the fourteen items, in which girls were more likely to succeed conditional on the ability level. This finding may have various explanations and implications. For instance, the fact that a considerable number of items were subject to gender DIF indicates that the ICT literacy, knowledge-lean test functioned differently across gender. On the one hand, this result may be considered a threat to the construct validity and test fairness of the test (Senkbeil, Ihme, & Wittwer, 2013), because of different success probabilities for boys and girls of the same level of ability (AERA, APA, & NCME, 2014; Millsap, 2011). On the other hand, gender DIF may have substantive reasons the eight items have flagged. All the 14 items are based on the competence aims from the curriculum, and we expected, therefore, that all students had equally good prospects of success. Hence, the fact that gender DIF occurred, although students may have been exposed to comparable learning opportunities – as specified in the Norwegian curriculum – may point to the existence of different response styles or the understanding of the test questions as alternative explanations. Moreover, differences in self-efficacy and ICT attitudes may be alternative explanatory variables (e.g., Broos, 2005; Sieverding & Koch, 2009). The current study is unfortunately not able

to explain *why* gender DIF occurred on empirical ground; yet, it provides insights into *how* and *where* gender DIF occurred. We believe that in-depth studies on the learning opportunities and students' response processes can supplement our findings in order to explain DIF; in this respect, explanatory item response modeling approaches provide valuable tools to quantify the effects of potential explanatory variables (De Boeck & Wilson, 2004). Given that the ICT literacy test was knowledge-lean, differences in both learning opportunities and individual levels of acquired knowledge are candidates for such variables. In fact, gender DIF occurred particularly for questions that tapped students' knowledge about terms or devices on the one hand (e.g., Q1 & Q2), and the credibility of digital information or processes (e.g., transparency and reproducibility of searches; Q8 & Q9) on the other hand. Considering the diversity of these topics and questions, no clear picture of potential substantive reasons could be drawn based on item contents. We suspect that the distinction between technology- and information-related aspects may explain part of the DIF variance across gender (Senkbeil et al., 2013). DIF analyses of single items can be helpful to detect "gender-biased" items in the test that were not discovered in the more global, multi-group measurement invariance test (e.g., Reynolds et al., 2008; Sideridis, Tsaousis, & Al-harbi, 2015). This discrepancy – the result that multi-group measurement invariance testing pointed to scalar invariance which was contradicted by the result of differential item functioning using the MIMIC-DIF approach – might have both methodological and substantive explanations. From a methodological point of view, the two approaches – multi-group CFA invariance testing and MIMIC-DIF modeling – differ in their capabilities to detect deviations of items from invariance (Bauer, 2016). Whereas the former tests the equivalence of a measure based on the overall measurement model (i.e., construct level), the latter tests hypotheses on DIF for specific items or item sets (i.e., item level). The multi-group CFA approach might

therefore camouflage item-specific DIF. The two approaches may thus supplement each other, and it is generally recommended to take them both (e.g., Marsh, Tracey, & Craven, 2006; Steinmetz et al., 2009). It must also be noted that multi-group CFA invariance testing is not per se unable to detect item-specific DIF; the sensitivity of the associated procedures (e.g., comparing the fit between models with and without equality constraints on specific item parameters) is however lower than in the MIMIC-DIF approach (Bauer, 2016). From a substantive point of view, however, our study reaches its limits; the current data do not reveal potential substantive reasons of students' response behavior or determinants thereof. Thus, we cannot be certain about the extent to which this discrepancy is due to substantive reasons that are linked to students' response behavior, their perceptions of the items or technology in general, or their competence beliefs. We thus encourage studies using think-aloud protocols to retrieve possible reasons and more substantive explanations that substantiate the findings from our study.

The third research question addressed the gender differences on the ICT literacy test. Comparison of the ICT literacy scores from boys and girls revealed that girls overall performed better than boys. This finding is consistent with other research on gender differences in students' academic performance (Voyer & Voyer, 2014). Moreover, it seems as if the commonly identified "gender gap" in ICT literacy in favour of boys could not be confirmed in the present study (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014; Genlott & Grönlund, 2016; Madigan, Goodfellow, & Stone, 2007). It is noteworthy that the proposed ICT literacy test did not only show gender differences in the overall performance (i.e., construct level) but in the probability of solving eight out of fourteen items correctly conditional on the ability (i.e., item level). To the best of our knowledge, we perceived that a considerable number of studies on gender differences in ICT-related

constructs focused on the construct level only, neglecting the potential existence of gender DIF at the same time. Hence, we encourage researchers to conduct DIF analyses together with analyses of performance differences to ensure that mean differences in performance are meaningful (i.e., they are not caused by the differential functioning of items; e.g., Millsap, 2011; Scherer & Siddiq, 2015).

Although the current study was conducted only within the Norwegian context, the ICT literacy test might well be applicable in other countries or educational systems, as it taps sub-categories of ICT literacy that are commonly used in conceptual frameworks of this construct (Ferrari, 2012; Hatlevik, 2017). At the same time, we are aware that empirical evidence is needed to support this claim, particularly with respect to the measurement invariance of the test across gender, countries, and educational systems. In fact, it is an important objective of Norwegian education to provide equal opportunities to all children and adolescents who attend school. It is a goal to avoid that socio-economic background, ethnicity or gender should have importance for the opportunities the school provides for the students. The schools are therefore expected to develop and apply initiatives and actions to compensate and equalize any differences and inequalities. In this way, it is important with research related to analysis of the fairness of tests and examining if the test works in the same way across gender.

5. Conclusions

This paper focused on the measurement invariance and differential item functioning of a Norwegian ICT literacy test. Multi-group confirmatory factor analysis showed that the test exhibited a sufficient level of measurement invariance across gender, which enables researchers to conduct meaningful mean comparisons. In the current study, these mean comparisons pointed to higher ICT literacy of girls in

comparison to boys. Yet, at the same time, differential item functioning existed for some items. This result indicated only minor deviations from perfect comparability across gender and therefore provides evidence that might be used to craft a validity argument for the ICT literacy test. Overall, the present study highlights the importance of testing for both measurement invariance and differential item functioning when examining gender differences in ICT literacy.

Figures

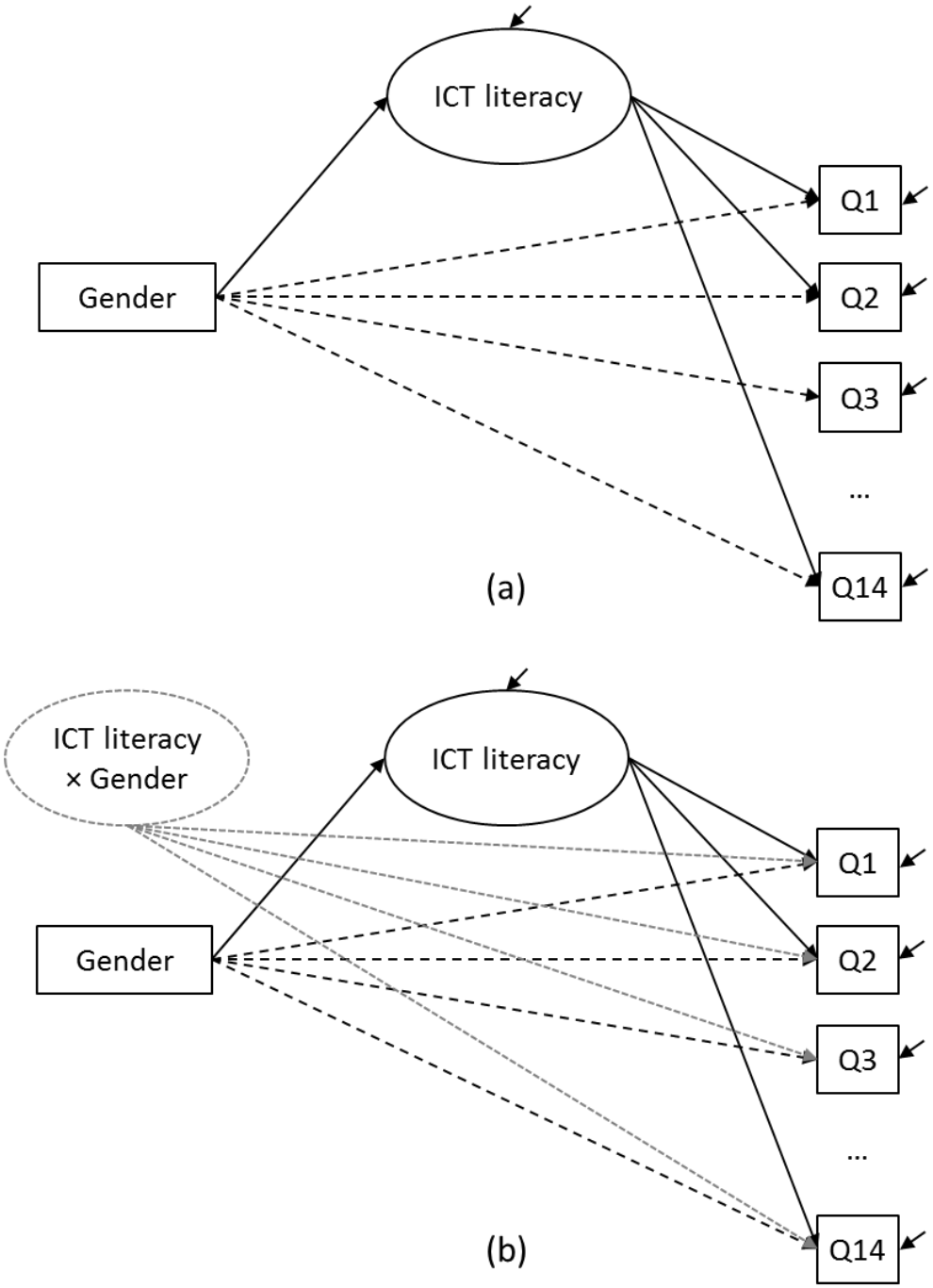

Figure 1. (a) MIMIC model to test for *uniform* DIF in the ICT literacy items Q_i; (b) MIMIC model to test for *non-uniform* DIF in the ICT literacy items Q_i.

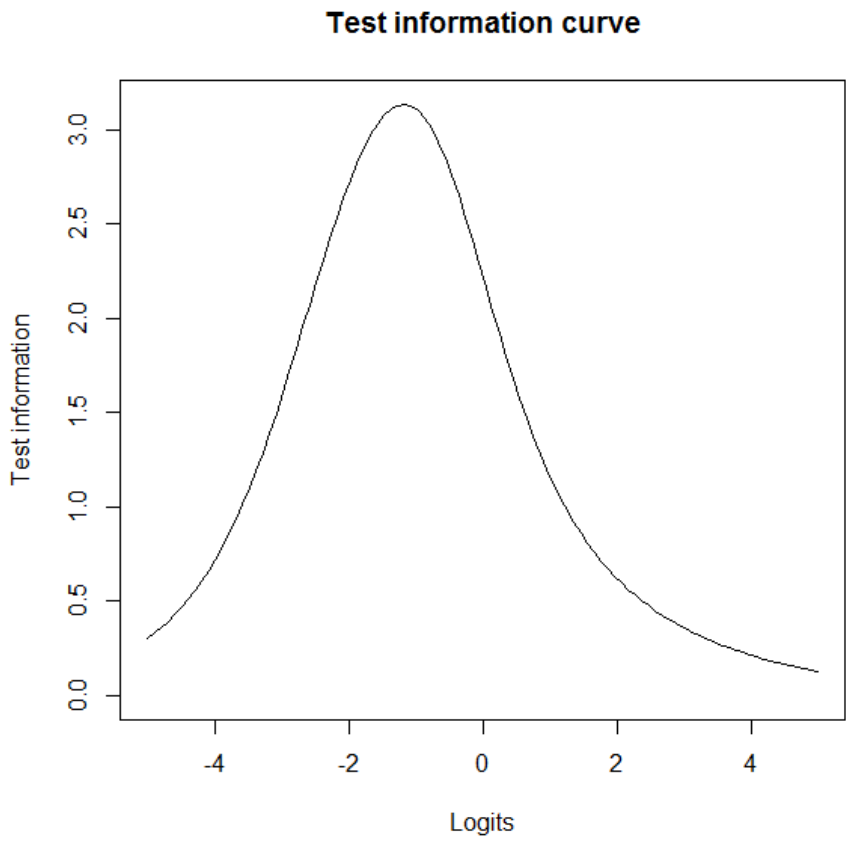

Figure 2. Test information curve of the ICT literacy test under the MLR estimation.

References

Acock, Alan C. (2012). *A gentle introduction to Stata* (revised 3rd edition). College Station, TX: Stata Press.

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association (AERA).

Aesaert, K., van Nijlen, D., Vanderlinde, R., Tondeur, J., Devlieger, I. and van Braak, J. (2015). The contribution of pupil, classroom and school level characteristics to primary school pupils' ICT competences: A performance-based approach, *Computers & Education, 87*, 55-69. dx.doi.org/10.1016/j.compedu.2015.03.014

Ainley, J., Fraillon, J., & Freeman, C. (2007). *National assessment program: ICT literacy years 6 & 10 report 2005*. Australia: MCEETYA.

Asparouhov, T., & Muthén, B. (2010). Weighted least squares estimation with missing data. Los Angeles, CA: Muthén & Muthén. Retrieved from (23 May 2017): https://www.statmodel.com/download/GstrucMissingRevision.pdf

Beauducel, A., & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186-203. doi:10.1207/s15328007sem1302_2

Binkley, M., Erstad, E., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw, and E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht, The Netherlands: Springer. doi:10.1007/978-94-007-2324-5

Bauer, D.J. (2016). A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychological Methods*. Advance Online Publication: doi: 10.1037/met0000077

Broos, A. (2005). Gender and Information and Communication Technologies (ICT) Anxiety: Male Self-Assurance and Female Hesitation. *CyberPsychology & Behavior*, *8*(1), 21-31. doi:10.1089/cpb.2005.8.21

Brown, T.A. (2006). Confirmatory Factor Analysis for Applied Science. London : The Guildford Press

Calvani, A., Fini, A., Ranieri, M., & Picci, P. (2012). Are young generations in secondary school digitally competent? A study on Italian teenagers. *Computers & Education*, *58*, 797-807.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC Methods for Detecting DIF Among Multiple Groups: Exploring a New Sequential-Free Baseline Procedure. *Applied Psychological Measurement*, *40*(7), 486-499. doi:10.1177/0146621616659738

Claro, M., Preiss, D. D., San Martín, E., Jara, I., Hinostroza, J. E., Valenzuela, S., & Nussbaum, M. (2012). Assessment of 21st Century ICT skills in Chile: Test design and results from high school level students. *Computers & Education*, *59*(3), 1042-1053. doi:10.1016/j.compedu.2012.04.004

Crocker L. & Aligna J. (2006). *Introduction to classical and modern test theory.* Independence, KY: Wadsworth Publishing Company.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* London: Guilford Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Theory Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer

Science+Business Media.

Dimitrov, D.M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, 43, pp. 121–149.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*(11), 2417-2425. doi:10.1037/a0037996

Elosua, P., & Mujika, J. (2015). Partial scalar invariance and observed differences across gender in a reasoning test battery. *Psicothema, 27*(3), 296-302. doi:10.7334/psicothema2014.282

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford Press.

Ferrari, A. (2012). *Digital Competence in Practice: An Analysis of Frameworks.* Report EUR25351EN. Luxembourg: Publications Office of the European Union.

Ferrari, A. (2013). *DIGCOMP: A framework for developing and understanding digital competence in Europe*. Report EUR26036EN. Luxembourg: Publications Office of the European Union.

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.* Heidelberg, New York, Dordrecht, London: Springer International Publishing.

Genlott, A. A., & Grönlund, Å. (2016). Closing the gaps - Improving literacy and mathematics by ict-enhanced collaboration. *Computers & Education*, *99*, 68-80. doi:10.1016/j.compedu.2016.04.004

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st

century. *Educational Research Review, 13*, 74-83. doi:10.1016/j.edurev.2014.10.002

Gui, M., & Argentin, G. (2011). Digital skills of internet natives: Different forms of internet literacy in a random sample of northern Italian high school students. *New Media & Society, 13*(6), 963–980.

Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly*, *87*(2), 432–448. doi:10.1111/j.1540-6237.2006.00389.x

Hatlevik O. E., & Christophersen, K-A. (2013). Digital competence at the beginning of upper secondary school: Identifying factors explaining digital inclusion. *Computers & Education, 63*, 240-247.

Hatlevik, O. E. (2017). Examining the Relationship between Teachers' Self-Efficacy, their Digital Competence, Strategies to Evaluate Information, and use of ICT at School, *Scandinavian Journal of Educational Research*, DOI: 10.1080/00313831.2016.1172501

Hatlevik, O. E., Throndsen, I., & Ottestad, G. (2015). Predictors of Digital Comp. in 7th Grade: Students' Motivation, Family Background, and Culture for Professional Development in Schools. *Journal of Computer-Assisted Learning, 31*(3), 220-231. 10.1111/jcal.12065

Hatlevik, O. E., Gudmundsdottir, G. B., & Loi, M. (2015). Digital diversity among upper secondary students: A multilevel analysis of the relationship between cultural capital, self-efficacy, strategic use of information and digital competence. *Computer & Education*, *81*, 345-353.

Hatlevik, O. E., Throndsen, I. & Loi, M. (2015). Kartlegging av digitale ferdigheter. In: O. E. Hatlevik & I. Throndsen (Eds.), *Læring av IKT. Elevenes digitale ferdigheter og*

bruk av IKT i ICILS 2013 (pp. 49-78). Oslo: Universitetsforlaget.

Hohlfeld, T. N., Ritzhaupt, A. D., & Barron, A. E. (2013). Are gender differences in perceived and demonstrated technology literacy significant? It depends on the model. *Educational Technology Research and Development, 61*(4), 639-663. doi:10.1007/s11423-013-9304-7

Holland, P.W., & Wainer, H. (1993). Differential Item Functioning. Hillsdale. NJ: Lawrence Erlbaum.Imhof, M., Vollmeyer, R., & Beierlein, C. (2007). Computer use and the gender gap: The issue of access, use, motivation, and performance. *Computers in Human Behavior, 23*(6), 2823-2837. doi:10.1016/j.chb.2006.05.007

Kim, J. M., & Lee, W.G. (2011). Meanings of criteria and norms: Analyses and comparisons of ICT literacy competencies of middle school students. *Computers & Education*, *64*, 81-94. doi:10.1016/j.compedu.2012.12.018

Kim, H. S., Kil, H. J., & Shin, A. (2014). An analysis of variables affecting the ICT literacy level of Korean elementary school students. *Computers & Education, 77*, 29-38.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York and London: Guilford Press.

Koch, S. C., Müller, S. M., & Sieverding, M. (2008). Women and computers. Effects of stereotype threat on attribution of failure. *Computers & Education*, *51*(4), 1795-1803. doi:10.1016/j.compedu.2008.05.007

Lau, W. W. F., & Yuen, A. H. K. (2015). Factorial invariance across gender of a perceived ICT literacy scale. *Learning and Individual Differences*, *41*, 79-85.

Lennon, M., Kirsch, I., von Davier, M., Wagner, M., & Yamamoto, K. (2003). *Feasibility study for the PISA ICT literacy assessment: Report to network A.* Paris: OECD Publishing.

Litt, E. (2013). Measuring users' internet skills: A review of past assessments and a look

toward the future. *New Media & Society, 15*(4), 612-630.

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the Source of Misfit in Item Response Theory Models. *Multivariate Behavioral Research*, *49*(4), 354-371. doi:10.1080/00273171.2014.910744

Madigan, E. M., Goodfellow, M., & Stone, J. A. (2007). Gender, perceptions, and reality: technological literacy among first-year students. *ACM SIGCSE Bulletin*, *39*(1), 410-414.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics* (pp. 275-340). Mahwah, NJ: Lawrence Erlbaum.

Martin, A., & Grudziecki, J. (2006). DigEuLit: Concepts and Tools for Digital Literacy Development. *Innovation in Teaching and Learning in Information and Computer Sciences, 5*(4), 1-19. doi:10.11120/ital.2006.05040249

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320-341. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Tracey, D. K., & Craven, R. G. (2006). Multidimensional Self-Concept Structure for Preadolescents with Mild Intellectual Disabilities. *Educational and Psychological Measurement, 66*(5), 795-818. doi:10.1177/0013164405285910

Martin, A., & Grudziecki, J. (2006). DigEuLit: Concepts and tools for digital literacy development. *ITALICS, Innovation in Teaching and Learning in Information and Computer Sciences*, *5*(4), 249-267.

McAlpine, M. (2001). *A summary of methods of item analysis. Bluepaper Number 2.*

Glasgow: University of Glasgow. Retrieved from (11 May 2017) http://caacentre.lboro.ac.uk/dldocs/BP2final.pdf

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568-592. doi:10.1037/0021-9010.93.3.568

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, B., & Muthén, L. (1998-2015). *Mplus (Version 7.3)*. Los Angeles, CA: Muthén & Muthén.

Norwegian Directorate for Education and Training. (2012). *Framework for Basic Skills.* Oslo: The Norwegian Directorate for Education and Training.

Organisation of Economic Co-operation and Development (OECD). (2015). *Students, computers and learning: Making the connection*. Paris: OECD Publishing. doi:10.1787/9789264239555-en

Padilla-Meléndez, A., del Aguila-Obra, A. R., & Garrido-Moreno, A. (2013). Perceived playfulness, gender differences and technology acceptance model in a blended learning scenario. *Computers & Education, 63*, 306-317. doi:10.1016/j.compedu.2012.12.014

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist, 51*(1), 59-81. doi:10.1080/00461520.2016.1145550

Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence, 36*(3), 236-260. doi: 10.1016/j.intell.2007.06.003

Rubio, M. A., Romero-Zaliz, R., Mañoso, C., & de Madrid, A. P. (2015). Closing the gender gap in an introductory programming course. *Computers & Education, 82*, 409-420. doi:10.1016/j.compedu.2014.12.003

Scherer, R., & Siddiq, F. (2015). Revisiting teachers' computer self-efficacy: A differentiated view on gender differences. *Computers in Human Behavior*, *53*, 48-57. doi:10.1016/j.chb.2015.06.038

Schroeders, U. & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension across Test Media. *Educational and Psychological Measurement*, *71*(5), 849-869. doi:10.1177/0013164410391468

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The Test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity/Test zur Erfassung technologischer und informationsbezogener Literacy (TILT) im Nationalen Bildungspanel: Entwicklung, empirische Überprüfung und Validitätshinweise. *Journal for Educational Research Online*, *5*(2), 139-161.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I. and Scherer, R. (2016). Taking a future perspective by learning from the past. A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review*, *19*, 58-84.

Sideridis, G. D., Tsaousis, I., & Al-harbi, K. A. (2015). Multi-Population Invariance With Dichotomous Measures. *Journal of Psychoeducational Assessment, 33*(6), 568-584. doi:10.1177/0734282914567871

Sieverding, M., & Koch, S. C. (2009). (Self-)Evaluation of computer competence: How gender matters. *Computers & Education*, *52*(3), 696-701. doi:10.1016/j.compedu.2008.11.016

Silva-Maceda, G., Arjona-Villicaña, P. D., & Castillo-Barrera, F. E. (2016). More Time or Better Tools? A Large-Scale Retrospective Comparison of Pedagogical Approaches to Teach Programming. *IEEE Transactions on Education, 59*(4), 274-281. doi:10.1109/TE.2016.2535207

Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education*, *54*(5), 627-650. doi:10.1007/s11159-008-9105-0

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity, 43*(4), 599-616. doi:10.1007/s11135-007-9143-x

Teo, T. (2015). Comparing pre-service and in-service teachers' acceptance of technology: Assessment of measurement invariance and latent mean differences. *Computers & Education, 83*, 22-31. doi:10.1016/j.compedu.2014.11.015

Teo, T., Lee, C. B., Chai, C. S., & Wong, S. L. (2009). Assessing the intention to use technology among pre-service teachers in Singapore and Malaysia: A multigroup invariance analysis of the Technology Acceptance Model (TAM). *Computers & Education, 53*(3), 1000-1009. doi:10.1016/j.compedu.2009.05.017

Tømte, C., & Hatlevik, O. E. (2011). Gender-differences in Self-efficacy ICT related to various ICT-user profiles in Finland and Norway. How do self-efficacy, gender and ICT-user profiles relate to findings from PISA 2006. *Computers & Education, 57*(1), 1416-1424. doi:10.1016/j.compedu.2010.12.011

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Choosing between Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance.

Frontiers in Psychology, 4. doi:10.3389/fpsyg.2013.00770

Van Deursen, A. (2012). Internet skill-related problems in accessing online health information. *International Journal of Medical Informatics, 81*(1), 61-72.

Van Deursen, A., & Van Dijk, J. (2009). Improving digital skills for the use of online public information and services. *Government Information Quarterly, 26*(2), 333-340.

Van Deursen, A., Van Dijk, J., & Peters, O. (2011). Rethinking Internet skills. The contribution of gender, age, education, Internet experience, and hours online to medium- and content-related Internet skills, *Poetics*, *39*, 125-144.

Volman, M., van Eck, E., Heemskerk, I., & Kuiper, E. (2005). New technologies, new differences. Gender and ethnic differences in pupils' use of ICT in primary and secondary education. *Computers & Education, 45*(1), 35-55. doi:10.1016/j.compedu.2004.03.001

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin*, *140*(4), 1174-1204. doi:10.1037/a0036620

Whitley Jr., B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior*, *13*(1), 1-22. doi:10.1016/S0747-5632(96)00026-X

Woods, C. M. (2009). Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behavioral Research, 44*(1), 1-27. doi:10.1080/00273170802620121

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment*, *31*(4), 320-338. doi:10.1007/s10862-008-9118-9

Yang, J. H. (2012). Effects of high school ICT activities on students' digital literacy in

Korea. *Journal of Educational Technology*, *28*(2), 347–369.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Dissertation submitted to the University of California, LA.

Tables

Table 1

Percentage correct, standardized factor loadings and standard error of the ICT literacy items

Item label	ICT literacy items	% correct	Standardized factor loading (*SE*)
Q1	Which digital device can you use to measure up a 4-km nature trail?	82	0.51 (0.05)**
Q2	What is the main difference between a digital map and a map on paper?	84	0.37 (0.06)**
Q3	Which statement about Wikipedia is incorrect?	37	0.31(0.05)**
Q4	You believe there should be no homework for school students. How can you work for this matter?	52	0.36 (0.05)**
Q5	You will find out who is the host of the "Eurovision Song Contest" on the TV channel NRK. Which keyword(s) should you use?	85	0.62 (0.05)**
Q6	Is it right for you to write false things about others online?	91	0.54 (0.06)**
Q7	You have found a poem on the Internet that you will use in a task. What should you do?	67	0.70 (0.04)**
Q8	Can you rely on information from Wikipedia?	79	0.68 (0.04)**
Q9	Can people that you do not know identify the websites you have visited and your online search words?	57	0.53 (0.04)**
Q10	What is a Wiki?	38	0.33 (0.05)**
Q11	Which digital tool should you choose to communicate with the student in Japan during school hours?	55	0.43 (0.05)**
Q12	You must publish a project assignment on an open blog. Can you use any image from the Flickr Image Sharing Service (www.flickr.com) in this project task?	56	0.44 (0.05)**
Q13	You want to publish pictures of others. What is the best thing to do?	87	0.50 (0.05)**
Q14	Why should you refer to sources in project assignments?	26	0.42 (0.05)**

Note. ** $p < .01$

Table 2

Goodness-of-fit statistics and comparisons among multi-group invariance models

Measurement invariance models	χ^2	*df*	RMSEA [90% CI]	CFI	WRMR	$\Delta\chi^2(\Delta df)$	ΔRMSEA	ΔCFI
Model 1: Configural	224.58**	154	.032 [.022, .040]	.945	1.319	-	-	-
Model 2: Scalar	242.55**	166	.032 [.023, .040]	.940	1.412	19.3 (12), $p = .082$	.000	-.005
Model 3: Strict	293.53**	180	.037 [.029, .040]	.911	1.623	62.3 (26), $p < .001$	.005	-.034

Note. Models 2 and 3 were compared to the baseline Model 1. The χ^2 difference test was performed under the DIFFTEST option in M*Plus* (Muthén & Muthén, 1998-2015). ** $p < .01$

Table 3

*Chi-square statistics of the **uniform** MIMIC-DIF testing procedure for each of the 14 ICT literacy items, model comparisons, and DIF effects (**WLSMV estimator**)*

			Model comparisons		*Uniform DIF effects*			
Model	χ^2	*df*	$\Delta\chi^2(1)$	*p*	*B*	*SE*	*p*	β
Baseline	191.37**	90	-	-	-	-	-	-
Uniform DIF								
Q1	177.89**	89	13.27	0.000	-0.345	0.095	0.000	-0.172
Q2	180.14**	89	10.36	0.001	-0.321	0.100	0.001	-0.160
Q3	190.62**	89	1.04	0.307	-0.087	0.085	0.308	-0.043
Q4	184.73**	89	6.47	0.011	0.211	0.083	0.011	0.104
Q5	191.35**	89	0.03	0.872	0.016	0.097	0.871	0.008
Q6	189.21**	89	2.48	0.116	0.178	0.113	0.115	0.088
Q7	184.56**	89	7.89	0.005	0.228	0.081	0.005	0.112
Q8	191.07**	89	0.27	0.606	0.046	0.088	0.605	0.023
Q9	185.86**	89	5.69	0.017	-0.192	0.081	0.017	-0.096
Q10	187.12**	89	4.14	0.042	-0.170	0.083	0.042	-0.085
Q11	172.97**	89	17.44	0.000	0.348	0.083	0.000	0.170
Q12	190.90**	89	0.63	0.426	0.065	0.081	0.425	0.032
Q13	185.68**	89	5.88	0.015	-0.245	0.101	0.016	-0.122
Q14	190.95**	89	0.77	0.380	0.079	0.090	0.380	0.039

Note. Students' gender was coded as 0 (*boys*) and 1 (*girls*). All models assuming uniform DIF in single items Q_i were compared to the baseline model under the DIFFTEST option in M*Plus* (Muthén & Muthén, 1998-2015). Cells in grey indicate that this item exhibited uniform DIF.
** $p < .01$

Table 4

*Information criteria of the **uniform** MIMIC-DIF testing procedure for each of the 14 ICT literacy items, model comparisons, and DIF effects (**MLR estimator**)*

							Model comparisons		*Uniform DIF effects*				
Model	*LL*	Npar	SCF	AIC	BIC	aBIC	cLRT	*p*	*B*	*SE*	*p*	*OR*	β
Baseline	-6807.8	29	1.022	13673.6	13813.5	13721.4	-	-	-	-	-	-	-
Uniform DIF													
Q1	-6801.0	30	1.020	13662.0	13806.7	13711.4	14.20	0.000	-0.731	0.201	0.000	0.481	-0.175
Q2	-6802.6	30	1.021	13665.2	13809.9	13714.6	10.78	0.001	-0.624	0.195	0.001	0.536	-0.160
Q3	-6807.1	30	1.022	13674.2	13818.9	13723.7	1.37	0.241	-0.172	0.146	0.239	0.842	-0.046
Q4	-6804.9	30	1.021	13669.8	13814.5	13719.3	5.87	0.015	0.347	0.144	0.016	1.414	0.089
Q5	-6807.8	30	1.022	13675.6	13820.3	13725.0	0.00	1.000	0.005	0.225	0.983	1.005	0.001
Q6	-6806.5	30	1.023	13673.1	13817.8	13722.5	2.48	0.115	0.409	0.259	0.114	1.505	0.091
Q7	-6804.4	30	1.021	13668.7	13813.4	13718.1	7.01	0.008	0.498	0.190	0.009	1.645	0.102
Q8	-6807.8	30	1.022	13675.5	13820.2	13724.9	0.11	0.737	0.070	0.206	0.735	1.073	0.015
Q9	-6804.4	30	1.023	13668.7	13813.4	13718.1	6.61	0.010	-0.434	0.172	0.012	0.648	-0.101
Q10	-6805.5	30	1.020	13670.9	13815.6	13720.3	4.87	0.027	-0.324	0.149	0.030	0.724	-0.084
Q11	-6799.4	30	1.021	13658.8	13803.5	13708.2	17.11	0.000	0.600	0.146	0.000	1.821	0.151
Q12	-6807.6	30	1.022	13675.3	13820.0	13724.7	0.35	0.556	0.092	0.155	0.552	1.097	0.023
Q13	-6804.8	30	1.020	13669.6	13814.3	13719.0	6.28	0.012	-0.574	0.235	0.014	0.563	-0.126
Q14	-6807.6	30	1.022	13675.1	13819.8	13724.5	0.50	0.479	0.115	0.165	0.485	1.122	0.030

Note. Students' gender was coded as 0 (*boys*) and 1 (*girls*). All models assuming uniform DIF in single items Q_i were compared to the baseline model using the corrected Likelihood-ratio difference test (cLRT) with ΔNpar = 1 in M*Plus* (Muthén & Muthén, 1998-2015). Cells in grey indicate that this item exhibited uniform DIF. LL = Log-likelihood value, Npar = Number of parameters, SCF = Scaling correction factor, AIC = Akaike's information criterion, BIC = Bayesian information criterion, aBIC = sample size-adjusted BIC, OR = Odds ratio. ** $p < .01$

Table 5

*Information criteria of the **non-uniform** MIMIC-DIF testing procedure for each of the 14 ICT literacy items, model comparisons, and DIF effects (**MLR estimator**)*

							Model comparisons		*Non-uniform DIF effects*			
Model	*LL*	Npar	SCF	AIC	BIC	aBIC	cLRT	*p*	*B*	*SE*	*p*	*OR*
Non-Uniform DIF												
Q1	-6800.2	31	1.020	13662.4	13811.9	13713.4	1.62	0.203	-0.357	0.285	0.210	0.699
Q2	-6802.1	31	1.023	13666.1	13815.7	13717.2	0.95	0.331	-0.250	0.259	0.335	0.779
Q3	-6807.0	31	1.026	13675.9	13825.5	13727.0	0.24	0.623	0.100	0.204	0.626	1.105
Q4	-6804.8	31	1.020	13671.6	13821.1	13722.7	0.21	0.643	-0.089	0.194	0.647	0.915
Q5	-6806.7	31	1.023	13675.4	13824.9	13726.5	2.04	0.154	0.537	0.390	0.169	1.711
Q6	-6806.2	31	1.025	13674.4	13823.9	13725.4	0.63	0.426	0.311	0.397	0.433	1.365
Q7	-6803.9	31	1.021	13669.8	13819.3	13720.8	0.93	0.335	-0.351	0.369	0.342	0.704
Q8	-6807.6	31	1.026	13677.3	13826.8	13728.3	0.19	0.659	0.165	0.381	0.665	1.179
Q9	-6801.9	31	1.022	13665.7	13815.2	13716.8	5.06	0.024	-0.618	0.288	0.032	0.539
Q10	-6805.5	31	1.019	13672.9	13822.4	13724.0	0.01	0.905	0.020	0.203	0.920	1.021
Q11	-6797.6	31	1.019	13657.3	13806.8	13708.3	3.72	0.054	0.385	0.207	0.062	1.470
Q12	-6807.5	31	1.021	13677.1	13826.6	13728.1	0.18	0.668	0.100	0.235	0.671	1.105
Q13	-6803.3	31	1.019	13668.7	13818.2	13719.7	2.97	0.085	-0.623	0.373	0.095	0.536
Q14	-6807.6	31	1.025	13677.1	13826.6	13728.2	0.00	0.966	-0.011	0.237	0.964	0.989

Note. Students' gender was coded as 0 (*boys*) and 1 (*girls*). All models assuming non-uniform DIF in single items Q_i were compared to the corresponding uniform DIF models using the corrected Likelihood-ratio difference test (cLRT) with ΔNpar= 1 in M*Plus* (Muthén & Muthén, 1998-2015). Cells in grey indicate that this item exhibited non-uniform DIF. LL = Log-likelihood value, Npar = Number of parameters, SCF = Scaling correction factor, AIC = Akaike's information criterion, BIC = Bayesian information criterion, aBIC = sample size-adjusted BIC, OR = Odds ratio. ** $p < .01$