



# Learning in Digital Networks – ICT literacy: A novel assessment of students' 21st century skills



Fazilat Siddiq <sup>a, \*</sup>, Perman Gochyyev <sup>b</sup>, Mark Wilson <sup>b</sup>

<sup>a</sup> Department of Teacher Education and School Research (ILS), Faculty of Educational Sciences, University of Oslo, Norway

<sup>b</sup> Berkeley Evaluation and Assessment Research (BEAR) Center, University of California, Berkeley, USA

## ARTICLE INFO

### Article history:

Received 29 August 2016

Received in revised form 27 January 2017

Accepted 29 January 2017

Available online 16 February 2017

### Keywords:

Computer-based assessment

Education

ICT literacy

Learning in Digital Networks (LDN-ICT)

Multidimensional item response theory

21st century skills

## ABSTRACT

The present investigation aims to fill some of the gaps revealed in the literature regarding the limited access to more advanced and novel assessment instruments for measuring students' ICT literacy. In particular, this study outlines the adaption, further development, and validation of the Learning in Digital Networks—ICT literacy (LDN-ICT) test. The LDN-ICT test comprises an online performance-based assessment in which real-time student-student collaboration is facilitated through two different platforms (i.e., GoogleDocs and chat). The test attempts to measure students' ability in handling digital information, to communicate and collaborate during problem solving. The data are derived from 144 students in grade 9 analyzed using item response theory models (unidimensional and multidimensional Rasch models). The appropriateness of the models was evaluated by examining the item fit statistics. To gather validity evidence for the test, we investigated the differential item functioning of the individual items and correlations with other constructs (e.g., self-efficacy, collective efficacy, perceived usefulness and academic aspirations). Our results supported the hypothesized structure of LDN-ICT as comprising four dimensions. No significant differences across gender groups were identified. In support of existing research, we found positive relations to self-efficacy, academic aspirations, and socio-economic background. In sum, our results provide evidence for the reliability and validity of the test. Further refinements and the future use of the test are discussed.

© 2017 Elsevier Ltd. All rights reserved.

## 1. Introduction

The rapid global changes in information and communication technology (ICT), has affected the availability of technology and made it pervasive. The permeation of technology in society has forced changes in employment and education. The new skills needed for navigating education and the workplace in the current century have been labeled 21st century skills (Griffin, McGaw, & Care, 2012) and are characterized as being critical for functioning effectively in society (Dede, 2009; Griffin, McGaw, & Care, 2012; Partnership for 21st Century Skills [P21], 2012). A number of national and international frameworks have been outlined to systematize and define 21st century skills (Binkley et al., 2012; Ferrari, 2013; Fraillon, Schulz, & Ainley, 2013). Furthermore, 21st century skills have been embedded in the national curriculum of a large number of countries

\* Corresponding author. University of Oslo, Faculty of Educational Sciences, Department of Teacher Education and School Research (ILS), Postbox 1099 Blindern, 0851 Oslo, Norway.

E-mail addresses: [fazilatu@gmail.com](mailto:fazilatu@gmail.com), [fazilat.siddiq@ils.uio.no](mailto:fazilat.siddiq@ils.uio.no) (F. Siddiq).

(Ananiadou & Claro, 2009; Balanskat and Gertsch, 2010; Gordon et al., 2009), which necessitates the monitoring and measurement of the skills students should attain.

However, research has identified that the assessment of 21st century skills is lagging behind (Voogt & Roblin, 2012), and most scholars agree that current assessment instruments do not successfully measure 21st century competences and call for authentic and complex tasks (e.g., Darling-Hammond & Adamson, 2010; Pepper, 2011; Silva, 2008). Researchers have specifically pointed out a lack of instruments for measuring students' skills related to digital communication, collaboration, and problem solving (Quellmalz, 2009; Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016; Wilson, Scalise, & Gochyyev, 2015). Moreover, in a recent systematic review, insufficient reporting of reliability evidence and the validity argument of such tests has been pointed out as problematic (Siddiq et al., 2016). The researchers further argued that this limits the trustworthiness of the tests as well as the credibility of the test scores and their interpretations. In addition, deficient reporting of the psychometric properties of the tests may restrict the reuse of the tests in future studies and samples. Consequently, there is a need for valid and reliable assessment instruments that target students' competence in solving problems, communicating, and collaborating through digital channels.

Against this background, the present study investigates the psychometric properties of the translated and revised Learning in Digital Networks—ICT literacy (LDN-ICT) test, which aims to measure students' ICT literacy in authentic, collaborative digital environments while staying within the boundaries of an objective assessment in classrooms (for more information about the initial test, see Wilson & Scalise, 2015; Wilson et al., 2015). More specifically, on the basis of a Norwegian sample, we explore (1) the dimensionality of the construct, (2) differential item functioning (DIF) with respect to gender and socio-economic status (SES), and (3) relations to other constructs. The overarching aim of the study presented in this paper is to gather and present evidence of the quality of the instrument with reference to commonly formulated aspects of reliability and validity (AERA, APA, & NCME, 2014; Messick, 1995).

## 2. Theoretical framework

### 2.1. Defining 21st century skills

A number of initiatives over the last years have proposed definitions and outlined frameworks of 21st century skills. These definitions reflect the importance of 21st century skills for individuals and for society. The Organization for Economic Co-operation and Development (OECD) defined 21st century skills as “those skills and competencies young people will be required to have in order to be effective workers and citizens in the knowledge society of the 21st century” (Ananiadou & Claro, 2009, p. 8). The Partnership for 21st Century Skills (P21; 2010) chose slightly different wording, and defined 21st century skills as “the skills, knowledge and expertise students must master to succeed in work and life—a blend of content, knowledge, specific skills, expertise and literacies” (p. 8). The definition proposed by the Assessment and Teaching of 21st Century Skills (ATC21S) project is to a great extent along these lines and proposes that 21st century skills are underpinned by knowledge, skills, attitudes, values, and ethics — referred to as the KSAVE model (Binkley et al., 2012). Although the definitions of 21st century skills differ slightly and to some extent connote different skills and/or reflect different categorizations of the skills, Voogt and Roblin's (2012) analysis of eight 21st century skills frameworks showed that the skills of collaboration, communication, digital literacy, citizenship, problem solving, critical thinking, creativity, and productivity are mentioned in most of the frameworks. Further, they emphasize that, in general, most 21st century skills frameworks are consistent with each other (Voogt & Roblin, 2012). In the following section, we will provide an overview of the ATC21S project and introduce the 21st century skills framework the project developed. Moreover, the LDN-ICT test aims to conceptualize certain skillsets in the ATC21S framework.

### 2.2. ATC21S project and framework

The international ATC21S project was established in 2009 as a means to meet some of the challenges of the world economy due to the changes brought by ICT developments (Wilson et al., 2015). The project consisted of three ICT companies (Cisco, Intel, and Microsoft) and a broad group of academics from relevant disciplines (Griffin et al., 2012). Two published books describe the project and report the initial findings. The first book aims to describe the different phases of the project, starting from the initial work of defining 21st century skills, developing the frameworks, and scrutinizing the methodological issues regarding the assessment of more novel competencies (Griffin et al., 2012). As part of the project, two sample assessments were developed for measuring certain competences (e.g., ICT literacy in digital networks and collaborative problem solving) within the 21st century framework. The preliminary results and experiences from the development and pilot studies of the two assessments are described in the second book (Griffin & Care, 2015).

The ATC21S project identified 10 skillsets that comprise the 21st century framework. The skillsets were organized in four groupings, including ways of thinking (creativity and innovation; critical thinking, problem solving, decision making; learning to learn, metacognition); ways of working (communication; collaboration); tools for working (information literacy; ICT literacy); and living in the world (citizenship; life and career; personal and social responsibility) (Griffin et al., 2012). Hesse, Care, Buder, Sassenberg, and Griffin (2015) accentuate the multifaceted, multidimensional, and complex nature of each of the skillsets.

### 2.3. Framework underlying the LDN-ICT test

The LDN-ICT test is the main focus of this study. This test is targeted at measuring the following main three skillsets in the ATC21S framework: information literacy, ICT literacy, and personal and social responsibility. It has been argued that these skillsets “could be conflated into the ways in which students learn through social networks and social media” (Griffin & Care, 2015, p. 7). The skillsets were further detailed, and in-depth descriptions of the underlying concepts were outlined. The LDN-ICT test is made up of four strands, which are seen as interconnected in the activity of learning in networks,<sup>1</sup> as follows:

- Functioning as a Consumer in Networks (CiN)
- Functioning as a Producer in Networks (PiN)
- Participating in the Development of Social Capital through Networks (SCN)
- Participating in the Development of Intellectual Capital (i.e., collective intelligence) in Networks (ICN)

An overview of the strands, including the levels and related competences of each, are described in Table 1. The strand *Functioning as a Consumer in Networks (CiN)* involves obtaining, managing, and utilizing information and knowledge from shared digital resources and experts to benefit private and professional lives. *Functioning as a Producer in Networks (PiN)* involves creating, developing, organizing, and re-organizing information and knowledge to contribute to shared digital resources. Further, *Social Capital through Networks (SCN)* involves using, developing, moderating, leading, and brokering the connectivity within and between individuals and social groups to marshal collaborative action, build communities, maintain an awareness of opportunities, and integrate diverse perspectives at community, societal, and global levels. Lastly, the strand *Intellectual Capital through Networks (ICN)* involves understanding how tools, media, and social networks operate and using appropriate techniques through these resources to build collective intelligence and integrate new insights into personal understandings. The four strands—including the different levels within each—constitute the hypothesized construct maps of the LDN-ICT test (Table 1). At the lowest levels (i.e., CiN1, PiN1, SCN1, ICN1; Table 1) of each strand are the competencies that one would expect to see exhibited by a novice or beginner. At the top levels (i.e., CiN3, PiN3, SCN4, ICN4; Table 1) are the competencies that one would expect to see exhibited by an experienced person—someone highly literate in LDN-ICT. These construct maps are hierarchical and probabilistic. The notion *hierarchical* refers to the idea that a person who normally performs well on a task at a higher level would also be expected to perform well on a task at a lower level, while *probabilistic* refers to the idea that the maps represent the different probabilities that a given competence would be expected to be exhibited in a particular context rather than certainties that the competence would always be exhibited. Moreover, as shown in Fig. 1, the levels within each strand do not indicate the same fixed scale (e.g., the lower levels of one strand may be equivalent to the middle or even higher levels of other strands).

As described, the LDN-ICT test is constructed to measure how individuals operate, collaborate, and learn through the social media. Hesse et al. (2015) accentuate that due to the complex and multifaceted nature of the underlying framework of the LDN-ICT test, it consists of dimensions that describe both social and cognitive skill development. In addition, the skillsets are indirectly related to further skillsets and competences in the ATC21S framework (see Fig. 1.1 in Griffin & Care, 2015, p. 11). Moreover, each task in the test is constructed to measure more than one strand. Evidence from various empirical studies suggests that ICT literacy constructs are multidimensional. This is evident from studies on the assessment of students' ICT literacy (Aesaert, van Nijlen, Vanderlinde, & van Braak, 2014; Huggins, Ritzhaupt, & Dawson, 2014) and also on teachers' integration of and emphasis on ICT literacy (Siddiq, Scherer, & Tondeur, 2016). On the basis of evidence from previous research and *a priori* knowledge about the theoretical framework, we expect four dimensions (i.e., the strands CiN, PiN, SCN, and ICN) to better describe the construct in focus. Nevertheless, due to the interrelations between the strands, correlations between each are to be expected. First, they refer to the same overall concept of ICT literacy. Second, they may represent a sequence of processes; for instance, the strand CiN, which is related to handling digital information (e.g., obtaining, managing, and utilizing information) may be seen as a prerequisite for PiN, which is related to the further use of information (e.g., creating, developing, organizing, and re-organizing information for contributing to shared digital resources). The strand PiN also takes a step into the communication and collaboration aspects of SCN and ICN. Thus, the four dimensions appear to be interwoven and may lead to distinct but correlated dimensions.

In conclusion, the dimensionality of the construct is an important component of the evidence for the *internal validity* (Messick, 1995). Furthermore, to strengthen the argument for *external validity*, the relations between the measure and external variables may be investigated. This could add to the test's construct validity, which refers to the extent empirical data and substantive theory support the interpretations based on assessment outcomes (Messick, 1995). Hence, to investigate the external validity of the LDN-ICT test, we added a questionnaire that the students responded to after taking the test. In the following section, we review the research on relevant external variables.

<sup>1</sup> Please note that the descriptions of the underlying framework in the present study have been outlined in a previous publication on the pilot and initial validation of the LDN-ICT test (Wilson & Scalise, 2015). Hence, some parts of the theoretical framework section were adapted from this publication, in which the original test was described.

**Table 1**

Descriptions of the four strands of the ICT literacy framework.

Consumer in networks	Producer in networks	Social capital	Intellectual capital
Emerging consumer; CiN1	Emerging producer; PiN1	Emerging connector; SCN1	Emerging builder; ICN1
Performs basic tasks	Produces simple representations from templates	Participates in a social enterprise	Possesses knowledge of survey tools
Has no concept of credibility	Starts an identity	Is an observer or passive member of a social enterprise	Is able to make tags
Searches for pieces of information using common search engines (e.g., movie guides)	Uses a computer interface	Knows about social networks	Posts a question
Knows that tools exist for networking (e.g., Facebook)	Posts an artifact		
Conscious consumer; CiN2	Functional producer; PiN2	Functional connector; SCN2	Functional builder; ICN2
Selects appropriate tools and strategies (strategic competence)	Establishes and manages networks & communities	Encourages participation in and commitment to a social enterprise	Acknowledges multiple perspectives
Constructs targeted searches	Possesses awareness of planning for building attractive websites, blogs, games	Possesses awareness of multiple perspectives in social networks	Uses thoughtful organization of tags
Compiles information systematically	Organizes communication within social networks	Contributes to building social capital through a network	Understands mechanics of collecting and assembling data
Knows that credibility is an issue (webpages, people, networks)	Develops models based on established knowledge		Knows when to draw on collective intelligence
	Develops creative & expressive content artifacts		Shares representations
	Possesses awareness of security & safety issues (ethical and legal aspects)		
	Uses networking tools and styles for communication among people		
Discriminating consumer; CiN3	Creative producer; PiN3	Proficient connector; SCN3	Proficient builder; ICN3
Judges credibility of sources/people	Possesses team-situational awareness in process	Initiates opportunities for developing social capital through networks (e.g., support for development)	Understands and uses architecture of social media such as tagging, polling, role-playing, and modeling spaces to link to knowledge of experts in an area
Integrates information in coherent knowledge framework	Optimizes assembly of distributed contribution to products	Encourages multiple perspectives and supports diversity in networks (social brokerage skills)	Identifies signal versus noise in information
Conducts searches suited to personal circumstances	Extends advanced models (e.g., business models)		Interrogates data for meaning
Filters, evaluates, manages, organizes, and re-organizes information/people	Produces attractive digital products using multiple technologies/tools		Makes optimal choice of tools to access collective intelligence
Seeks expert knowledge (people through networks)	Chooses among technological options for producing digital products		Shares and reframes mental models (plasticity)
Selects optimal tools for tasks/topics			
		Visionary connector; SCN4	Visionary builder; ICN4
		Takes a cohesive leadership role in building a social enterprise	Questions existing architecture of social media and develops new architectures
		Reflects on experience in social capital development	Functions at the interfaces of architectures to embrace dialogue

Note: Consumer in Networks = CiN; Producer in Networks = PiN; Social Capital through Networks = SCN; Intellectual Capital through Networks = ICN.

#### 2.4. Relations to student beliefs (self-efficacy, collective efficacy, perceived usefulness, and academic aspirations)

Educators and researchers have been concerned with the identification of predictors of educational achievement (e.g., academic performance) to develop interventions and pedagogical strategies for students at risk of academic failure (Caprara, Vecchione, Alessandri, Gerbino, & Barbaranelli, 2011). Research on students' self-beliefs (e.g., self-efficacy, perceived

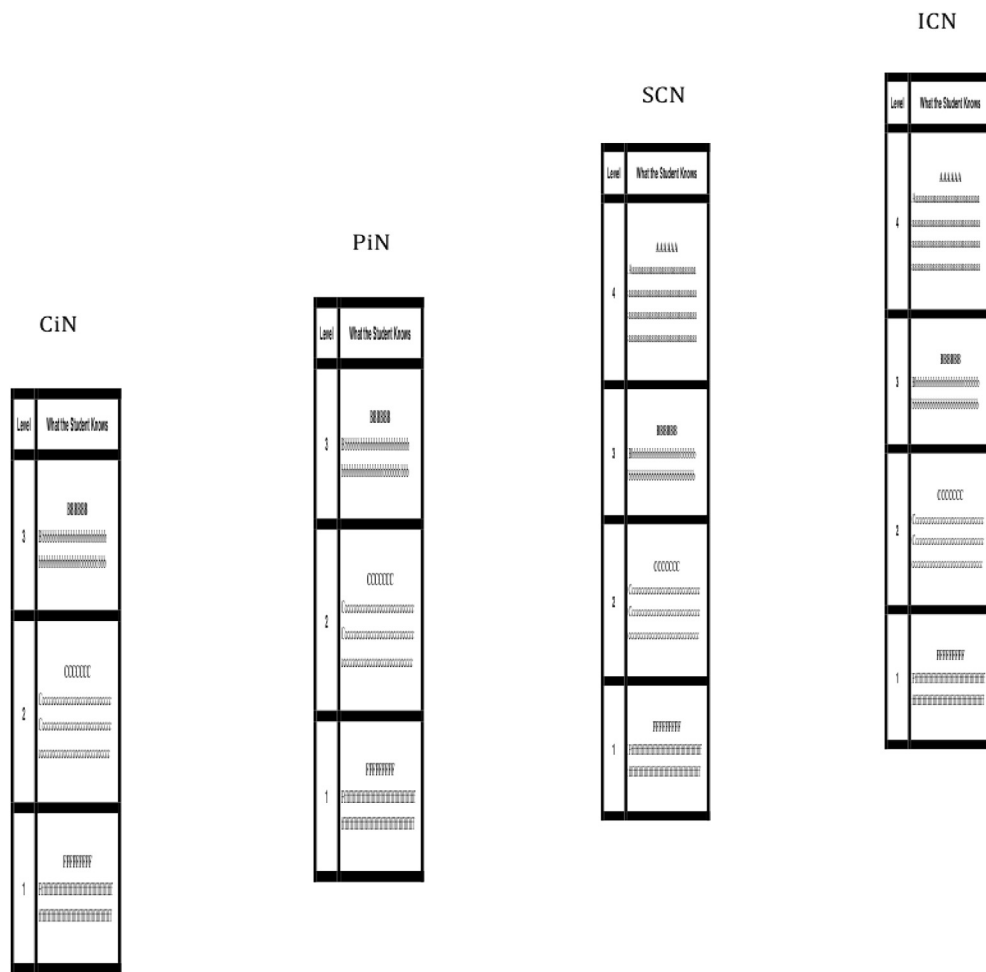


Fig. 1. The four strands of LDN-ICT represented as a four-part learning progression.

usefulness, and perceived ease of use) has shown that personal beliefs significantly affect academic performance (Caprara et al., 2008; Klassen, 2004; Martin, Montgomery, & Saphian, 2006; Pajares & Schunk, 2001; Robbins, Lauver, Davis, Langley, & Carlstrom, 2004; Marsh, Trautwein, Ludtke, & Baumert, 2006). The theory of planned behavior posits that behavior is affected by an individual's intention to perform it (Ajzen, 1991) and suggests that performance is a function of intentions and perceived behavioral control. Hence, following Ajzen (1991), we may expect positive relations between self-belief constructs and students' learning in digital networks.

**Self-efficacy.** Self-efficacy is a commonly used measure for studying the extent to which students believe in their own skills and has been studied as a model that can explain motivation and behavior. Bandura (1997) defined self-efficacy as "people's judgments of their capabilities to organize and execute courses of action required attaining designated types of performances" (p. 391). Research has revealed that self-efficacy is not a general (i.e., global) construct "but a differentiated set of self-beliefs linked to distinct realms of functioning" (Bandura, 2006, p. 307). Hence, the self-efficacy measure should reflect the specific knowledge domain or relevant aspects of the activity that is supposed to be measured. Compeau and Higgins (1995) extended the self-efficacy measure to include the technology perspective and defined computer self-efficacy as "a judgment of one's capability to use a computer" (p. 192). They argue that computer self-efficacy does not refer to simple component skills but rather judgments of ability to apply those skills to broader tasks.

Research on students' ICT literacy has documented positive and significant correlations between students' ICT literacy and self-efficacy (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014; Hohlfield, Ritzhaupt, & Barron, 2013). By means of investigating the external validity, we expect positive and high correlations between students' self-efficacy measure and their scores on the LDN-ICT test.

**Collective efficacy.** Collective efficacy (also commonly referred to as group efficacy) was defined by Bandura (1997) as a "group's shared beliefs in its conjoint capabilities to organize and execute the courses of action required to produce given levels of attainments" (p. 477). In contrast to self-efficacy, collective efficacy is concerned with the performance capability of



the group as a whole. The most commonly used approach to measure a group's perceived efficacy is concerned with the team members' appraisals of their group's capability operating as a whole (Stajkovic, Lee, & Nyberg, 2009). Furthermore, research has revealed positive correlations between collective efficacy and group performance (Bandura, 1997; Goddard, 2001; Greenlees, Graydon, & Maynard, 1999; Hodges & Carron, 1992; Peterson, Mitchell, Thompson, & Burr, 2000). Moreover, these findings have also been supported by meta-analysis studies (Gully, Incalcaterra, Joshi, & Beaubien, 2002; Stajkovic & Lee, 2001; Stajkovic et al., 2009). Thus, theory concerning how people work together within teams and other social units have accentuated collective efficacy as a “key social cognitive element that may help to explain how groups function together” (Lent, Schmidt, & Schmidt, 2006). Nevertheless, given the growing emphasis in education on students' collaboration and interaction in digital and social networks and collaborative problem-solving skills, collective efficacy has as of yet (to the best of our knowledge) not been applied in the context of 21st century skills assessment. Hence, as part of investigating the external validity of the LDN-ICT test, we are interested in the relations between collective efficacy and students' performance on the LDN-ICT test and expect positive and significant correlations between the two measures—especially the strands that include collaborative tasks (i.e., the strands SCN and ICN).

**Perceived usefulness.** Perceived usefulness of ICT is part of individuals' belief system and refers to the belief that ICT use will enhance their performance (Davis, 1989). Research on students' attitudes toward ICT has shown that students are more inclined to use ICT in academic settings when they perceive it as useful related to specific instances (Kirkwood & Price, 2005). Accordingly, Edmunds, Thorpe, and Conole (2010) revealed that the perceived usefulness of ICT and ICT usage and proficiency predicted students' attitudes toward ICT tools and their employment in different settings. On the other hand, a number of studies did not find significant relations between students' perceived usefulness of ICT and attributes such as learning (Keengwe, 2007) and course effectiveness (Venkatesh et al., 2012). Interestingly, most studies that included the perceived usefulness construct targeted older students (e.g., university or college students). Additionally, access to studies that have investigated the relations between students' ICT literacy and their perceived usefulness of ICT is scarce. Hence, as part of investigating the external validity of the LDN-ICT test, we investigate the relations between students' achievement on the LDN-ICT test and their perceived usefulness of ICT. Given the limited amount of available empirical research, it is not possible to state a clear hypothesis on the relationship between perceived usefulness and achievement. However, based on theoretical and common sense notions, we suggest that it is likely a positive one.

**Academic aspirations.** Students' aspirations and perceived academic success have been explored in several studies, and associations between students' academic expectations and achievement in several learning domains have been identified (Valentine, DuBois, & Cooper, 2004). Accordingly, in the International Computer and Information Literacy Study (ICILS), educational aspirations were measured by asking students to select the level of education they expected to attain. The results showed that, on average, computer and information literacy scores increased with levels of expected educational attainment (Fraillon et al., 2014). Hatlevik and Gudmundsdottir (2013) used the concept of academic aspirations in a slightly different way by asking lower secondary school students which study program (e.g., vocational training or general studies) they planned to attend in upper secondary school. Their results also revealed significant and positive relations between academic aspirations and students' achievement. In line with these findings, we expect positive and high correlations between students' LDN-ICT scores and academic aspirations.

## 2.5. Measurement invariance and differences across students' background characteristics (gender and SES)

In this section, we provide a brief description of measurement invariance, followed by an overview of the findings from previous studies on contextual factors (e.g., gender and SES) and their relations to ICT literacy. Such background factors have been identified as potential sources of variation in 21st century skills tests in precedent research.

**Measurement invariance.** We investigate the measurement invariance of two reasons. First, from a test development point of view, we aim to ensure that the test (i.e., the items or tasks) is not biased toward or against groups of students (Wilson, 2005) and thus provide evidence for the internal validity of the test. Second, for making valid comparisons across groups, it is critically important that the measurement invariance is sufficiently met; otherwise, mean comparisons are compromised (Millsap, 2011). Therefore, from a generalizability point of view, testing for measurement invariance provides an additional source of evidence for construct validity (Messick, 1995). Hence, we investigate the invariance (i.e., DIF) of the LDN-ICT test across gender and SES to study the comparability of the measurement model and potential differences in factor means. This is vital for drawing valid inferences on group differences.

**Gender.** Varying results on the relations between students' gender and their levels of ICT literacy have been reported (Kim, Kil, & Shin, 2014). Several studies have shown that female students score significantly higher than their male peers (Aesaert & van Braak, 2015; Baek et al., 2009; Hohlfeld et al., 2013; MCEETYA., 2007). In contrast, a number of studies have indicated that male students show more positive attitudes (Meelissen & Drent, 2008; Tsai, Tsai, & Hwang, 2010; Vekiri & Chronaki, 2008) and higher scores (Calvani, Fini, Ranieri, & Picci, 2012) than female students. Moreover, other studies on adolescents' ICT literacy and the effect of gender did not reveal significant differences (Claro et al., 2012; Durndell & Haag, 2002). These inconclusive findings on gender differences in ICT literacy assessment warrant a deeper look into whether gender differences occur for the LDN-ICT test, especially since this test aims to measure dimensions of ICT literacy (e.g., real-time student-student collaboration) that have scarcely been measured in previous research.

**SES.** Research on students' SES (e.g., parental educational level and occupation) have indicated that factors related to students' family background influence their ICT literacy outcome. Ainley and colleagues (ACARA., 2012) found that the

children of parents with low educational levels showed poorer ICT literacy proficiency than students who had parents with higher education (i.e., bachelor degree or above). Moreover, students from low-SES homes have been found to express lower ICT self-efficacy (Vekiri, 2010). Interestingly, Claro et al.'s study on the effects of economic, social, and cultural status on digital literacy versus reading and mathematics literacy showed that parents' level of education was the most relevant factor for explaining students score on the ICT test—more than for mathematics and reading achievement (Claro, Cabello, San Martin, & Nussbaum, 2015). Hence, we are interested in whether similar relations may occur between students' scores on the LDN-ICT test and their socio-economical background.

## 2.6. Item response theory (IRT)

IRT has been identified as a considerably useful methodology for test development and especially for investigating the reliability and validity of assessments (Hambleton & Jones, 1993). In contrast to classical test theory (CTT), in which the analyses are executed on the whole test and not the distinct items, the IRT approach focuses on individual items as building blocks of the test. In addition, it accentuates the underlying trait independent of the actual sample of respondents and items (Thomas, 2011; Wilson, 2005). Hence, the advantage of IRT is that the item parameter estimates are independent of the particular sample taken from a population of respondents, and the person estimates are independent of the specific sample of items administered to the person taken from a population of items (de Ayala, 2013). Moreover, both items and persons are given a score on the same scale simultaneously. Additionally, instead of a single aggregated measurement error such as in CTT, IRT gives a unique measurement error for each item and person, reflecting the fact that the reliability of a test depends on the unique interaction between the test material and the test taker (de Ayala, 2013).

Given the benefits of IRT, the Rasch measurement model was applied to address the main objectives of this study (i.e., to investigate the reliability and validity of the LDN-ICT assessment instrument). The Rasch measurement model was proposed by George Rasch (Rasch, 1960) and is the simplest model within the IRT framework; and when the items are shown to fit the model, this has certain advantages in interpretation. Moreover, data (each individual item) are tested for fit to the Rasch model, which facilitates a detailed examination of the internal construct validity of the scale, including properties such as reliability, ordering of categories, and invariance testing (Bond & Fox, 2007). Furthermore, the unidimensional Rasch model has been expanded due to the assumption of unidimensionality, which may be unrealistic in certain contexts. The multidimensional IRT model, which accounts for the possibility of a construct consisting of several factors, has proven to be useful. The multidimensional Rasch model and multidimensional IRT (MIRT) models in general were proposed as “categorical” variants of factor analysis methods and confirmatory factor analysis (CFA) in particular (McKinley & Reckase, 1983; Reckase, 1985). MIRT is sometimes referred to as full information factor analysis (Bock, Gibbons, & Muraki, 1988). This is because IRT models are fitted to the raw data directly rather than to summary statistics, such as polychoric correlations. CFA and multidimensional IRT (i.e., multidimensional Rasch) are both “confirmatory” methods aimed at confirming a hypothetical structure of the data. The multidimensional Rasch model (MRCML; Adams, Wilson, & Wang, 1997) is the simplest and the most parsimonious interpretation within the family of MIRT models. In comparison to the multidimensional two-parameter logistic model, this model assumes that the item loadings are fixed at unity.

Moreover, as previously discussed, the multidimensionality of the ICT literacy construct has been identified in several studies. However, to the best of our knowledge, multidimensional Rasch modeling has yet not been applied in this context.

## 2.7. The present study

Research on the assessment of students' ICT literacy has largely focused on students' competences in searching, retrieving, and evaluating digital information. Moreover, a gap between the content of 21st century skills frameworks and the operationalization of these has been identified (Siddiq et al., 2016). More specifically, tests which target student-student collaboration through digital channels and their problem-solving competences are scarce. However, initial research projects have been initiated to investigate such competences. The LDN-ICT is an example of a test measuring students' learning in digital networks; preliminary findings based on the pilot study indicated evidence for a reliable measure, and sound levels of internal structure validity were reported (Wilson & Scalise, 2015). The present study aims to further examine the reliability and validity evidence of the modified LDN-ICT literacy test and investigate the robustness of the test using data from Norwegian ninth grade students.

In particular, we address the following research questions:

1. To what extent can the evidence for the internal validity (e.g., reliability, item fit) of the LDN-ICT test be proved?
2. To what extent can the underlying conceptual framework with four dimensions be confirmed?
3. Does the LDN-ICT test provide an invariant measure across students' gender and socio-economic background?
4. To what extent can the evidence for the external validity (e.g., correlations to background variables; self-efficacy, collective efficacy, perceived usefulness of ICT, and academic aspirations) of the LDN-ICT test be provided?

### 3. Method

#### 3.1. Measures

The LDN-ICT test was translated, adapted, and revised to function in the Norwegian language style and educational system. Most of the tasks were kept close to the original version, yet the translation of some tasks required larger changes and replacements of some tasks. Before the main pilot, think-aloud protocols (also commonly referred to as *cognitive labs*) of 18 students were conducted in two rounds to investigate and improve the functionality, user experience, and content of the translated test.

The test consists of 39 items (i.e., 22 items in Arctic Trek and 17 items in Human Legacy), and some tasks were automatically scored while others were hand-scored. The hand-scored tasks comprised the more complex items that involved student interaction and were scored by following a scoring rubric developed in previous studies (e.g., Wilson & Scalise, 2015). Moreover, a number of items were dichotomous, while others were polytomous—reflecting a more fine-grained evaluation of students' performance on the items (Appendix A).

#### 3.2. The translated, adapted, and revised LDN-ICT test: two scenarios

The initial version of the LDN-ICT test was developed by the Berkeley Evaluation and Assessment Research (BEAR) Center at UC Berkeley.<sup>2</sup> The original test targeted students aged 11, 13, and 15. However, in the Norwegian version, only 15-year-old students were targeted, since experience from the pilot study showed that younger students struggled with some items in the test (Wilson & Scalise, 2015). Moreover, this is the age group that has been most frequently measured in other related studies (e.g., Arnseth, Hatlevik, Kløvstad, Kristiansen, & Ottestad, 2007; Fraillon et al., 2014), which may facilitate the comparison of different dimensions of students' 21st century skills.

Two different scenarios were chosen as the context in which tasks and items were placed along each of the four strands. Each scenario was designed to address more than one strand, but how the strands were represented in each scenario varied. The test-design consisting of two scenarios including several tasks has the benefit that test-users (e.g., teachers, researchers) may use the test in a flexible manner. Each scenario can for instance be used in classroom instruction or assessment individually or together. Moreover, the developers strived to use existing web-based tools when possible to lower both the costs and the familiarity of the tools (e.g., Google docs). A short description of each scenario is given below.

**Arctic Trek.** This scenario was constructed within the context of mathematics and natural science subjects, and the translated version was to a large degree kept close to the initial test, as described by Wilson and Scalise (2015). We were able to select Norwegian web sources about polar bears that were compatible with the webpages in English. However, while in the original test the Arctic Trek scenario was conceived as a collaboration contest or virtual treasure hunt, we labeled it only as collaboration task in the Norwegian test. Due to language and cultural differences, the Norwegian students would not have perceived the tasks in Arctic Trek as a collaboration contest or a treasure hunt. Except for the simplification of language and use of external web resources, the overall tasks, content, test-design, and required processes were retained.

**Sample tasks from Arctic Trek.** The students logged in to an online webpage. As shown in Fig. 2, the welcome screen consists of general information, the number of pages in the test, and an indication of which page the students were on as well as “Back” and “Next” buttons. These buttons are used to maneuver the test, while the rest of the page provides the information about the task the students are supposed to solve. The students' interaction with the test and the tasks is facilitated through links and web-sources, which open in new tabs in the Internet browser.

The students' goals while solving tasks in the Arctic Trek scenario were to find answers to six questions in collaboration with other students, and each student joined his/her team by following a link to a specific GoogleDoc. Once the team was assembled, it assigned roles to each team member (Fig. 3). The Team Notebook (i.e., a shared Google document pre-assigned to the usernames) facilitated the communication and collaboration between team members (Fig. 4).

First, the team received general information about the tasks (Fig. 5). Second, to become familiar with the structure of the tasks and the test environment, they started with a practice task, and the test-takers had to use the web resources listed in the right-hand panel to answer the questions (Fig. 6). If a student did not manage to solve the task, he or she could request a hint twice. The hints appear at the bottom of the screen (Fig. 6). If the hints did not sufficiently help the student to solve the problem, he/she was allowed request teacher assistance. In such cases, the teachers were asked to fill in information about the help they offered by hitting the “T” button in the bottom right-hand corner (Fig. 6).

**Human Legacy**<sup>3</sup>. This scenario was developed in the context of social science and arts subjects and framed as part of a poetry work unit in which students read and analyze well-known poems. As described by Wilson and Scalise (2015), attempts were made to keep the tasks close to authentic classroom situations. For instance, in a typical classroom context, the teacher might find it difficult to get the students to express the moods and meanings of a poem. Moreover, the students might wait to hear the teacher's understandings of the poem first and then agree with it.

<sup>2</sup> For more information, visit <http://bearcenter.berkeley.edu/>.

<sup>3</sup> Note that Wilson and Scalise (2015) labeled this scenario *Webspiration*. We chose the labeling *Human Legacy* instead due to the fact that *Webspiration* refers to the software used in the original test, which was not available for the translated test.



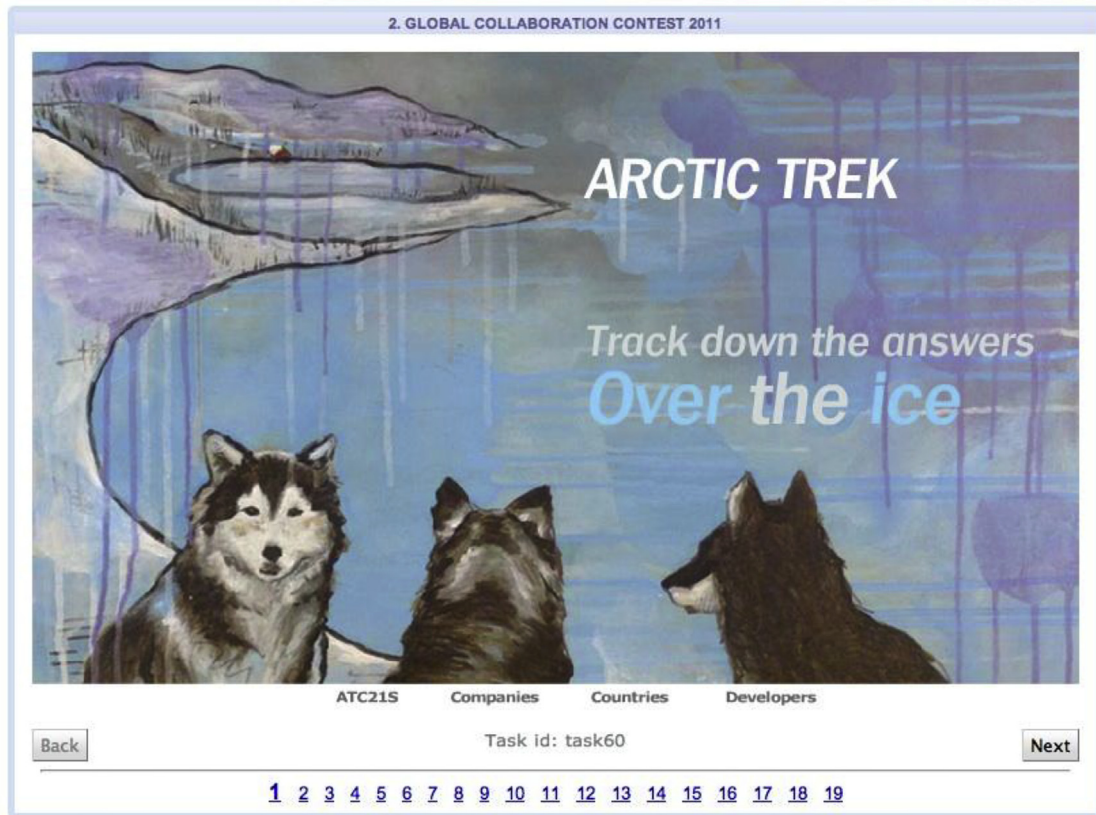


Fig. 2. The welcome screen from *Arctic Trek* scenario.

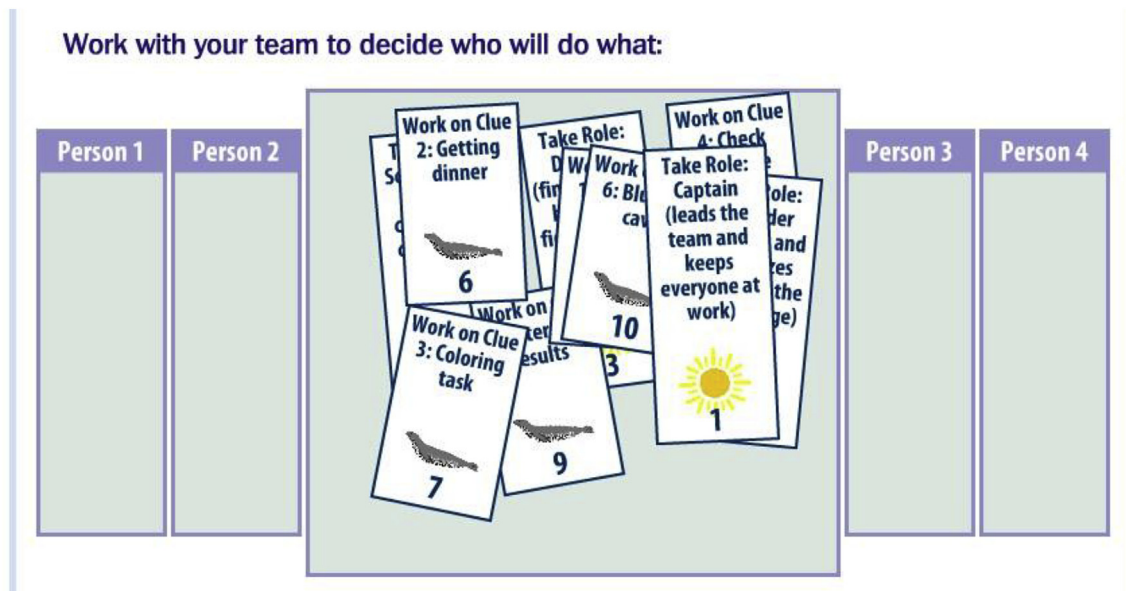


Fig. 3. Setting up the team roles and sharing tasks among team members. Note: The sample screens shown are from the English version of the test. The translated version (i.e., Norwegian) was kept similar except that the instructions were in Norwegian.

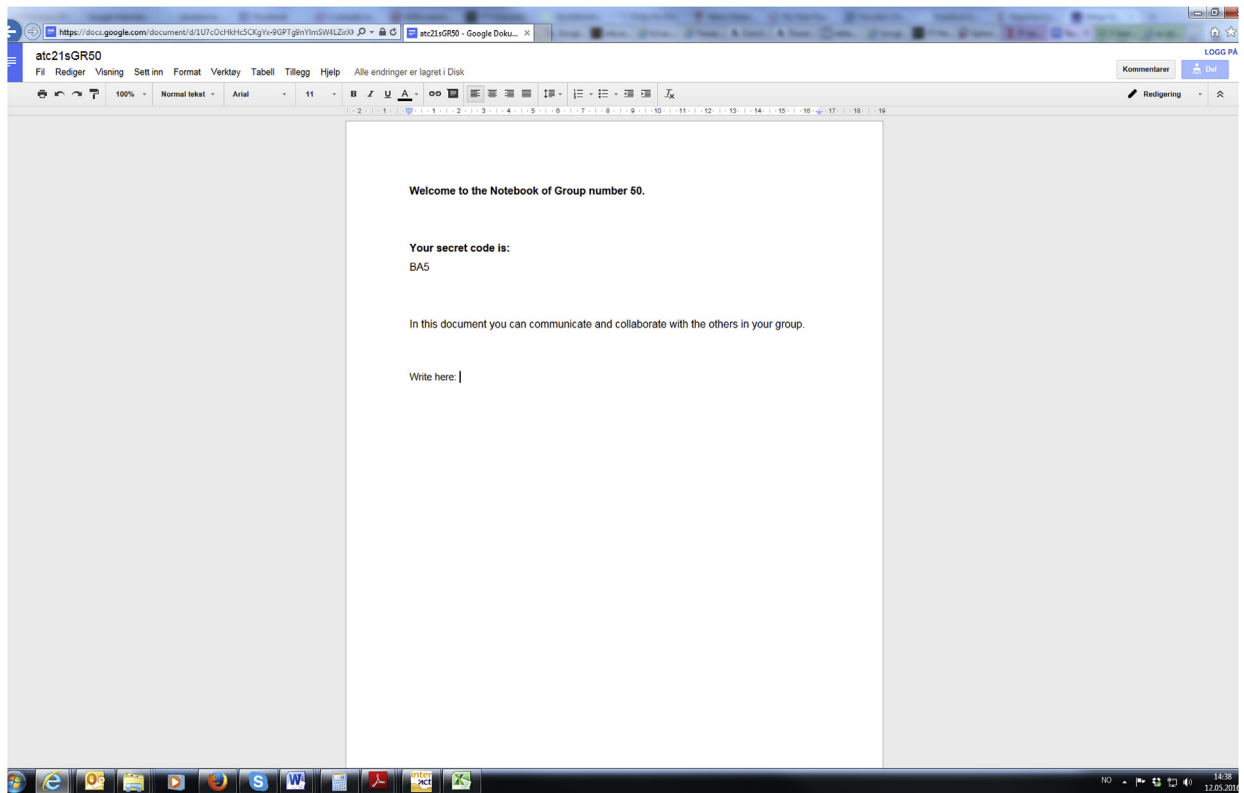


Fig. 4. Team notebook.

## ARCTIC TREK

### Collaboration task

For this collaboration, you work with your team and use clues to discover a series of 6 answers.

**HINT:**  
Here is how a clue works. The first part of the clue directs you to one of the web sites listed to the right. The rest of the clue guides you through the site to find the answer.

Track down the answers

### Over the ice

- [Finnish Arctic Club](#)
- [Polar Bear Population](#)
- [Polar Bear Map](#)
- [Land Animal Food](#)
- [Basic Computer Use](#)
- [Excel Spreadsheet](#)
- [Global Fishing](#)
- [Tagxedo](#)

Fig. 5. General information about the tasks.

The layout, design, and functionality of the two scenarios were kept similar for both scenarios. This was done to make it easier for the students to interact with the technology and to create the impression of one test, yet consisting of two parts.

**Sample tasks from Human Legacy.** In this scenario, some tasks in the translated version were slightly different from the original version. There are two particular reasons for the implemented changes. First, the online tool *Webspiration* was not available outside the United States and could consequently not be used in the translated version of the test. Second, no similar

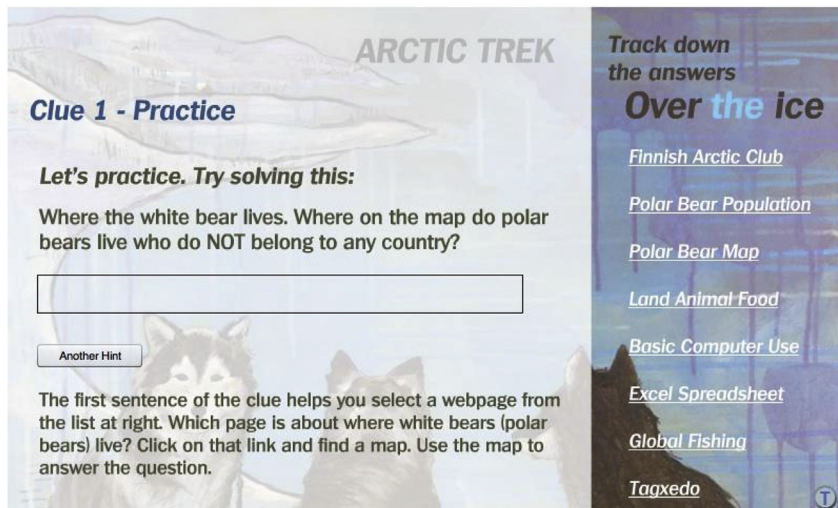


Fig. 6. A practice task, also showing the function of the hint button.

software within a reasonable price range could be identified. Consequently, a new task was developed within the Human Legacy scenario to capture students' collaborative problem-solving (ColPS) skills. We used the graphic online tool *Cosketch*<sup>4</sup> with embedded chat functionality for the ColPS task, in which three students were assigned to a team. To provide the same stimulus to all students and cause them reflect and develop their own ideas about the poem, they were given individual tasks before and after the ColPS task. First, the test-takers had to access an external web source to read a poem and watch a YouTube-video about it, followed by questions related to the poem. Second, the students were asked to create a mind map reflecting their understanding of the moods and meanings of the poem (Fig. 7). Subsequently, the students entered the ColPS task, in which they were asked to make a drawing of their interpretation of the poem together with their teammates (Fig. 8). The Cosketch software facilitated collaboration and communication by allowing the students to interact and discuss the poem through the embedded chat and sketch a drawing. The software allows for only one drawing—meaning that the functionality was equally shared between the students, and they worked and communicated in the same virtual room (Fig. 9). The remaining tasks in the Human Legacy scenario were individual and kept similar to those in the original test.

### 3.3. Background questionnaire

After taking the test, the students responded to a background questionnaire, which was developed to facilitate further investigations of the translated LDN-ICT tests' validity (e.g., external validity). The background questionnaire consists of constructs related to students' ICT self-efficacy, collective efficacy, perceived usefulness, academic aspirations, and background variables such as gender and SES. The constructs showed good reliability and were used to address RQ4 (Appendix B).

**ICT self-efficacy.** Students' ICT self-efficacy relates to their beliefs in using ICT for various learning purposes and was closely related to some of the content in the LDN-ICT test. It was measured using three items (e.g., *I am sure I know how to collaborate with other students by use of digital technology*), which the students rated on a six-point Likert scale from 1 (totally disagree) to 6 (totally agree). The results indicated that the instrument had acceptable reliability (Cronbach's alpha = 0.79).

**Collective efficacy.** Collective efficacy refers to the individuals' beliefs in their group's skills. Students' collective efficacy was measured by seven items (e.g., *I'm certain my team communicated adequately while working with the different collaboration tasks in the test*), which they rated on a six-point Likert scale ranging from 1 (totally disagree) to 6 (totally agree). Our results showed that the instrument had high reliability (Cronbach's alpha = 0.85).

**Perceived usefulness of ICT.** Students' perceived usefulness of ICT refers to the degree which they believe ICT would increase their performance in future learning and employment. This construct was assessed using six items (e.g., *Being able to collaborate digitally is important for working more efficiently*), which they rated on a six-point Likert scale ranging from 1 (totally disagree) to 6 (totally agree). Our results indicated an acceptable reliability of the instrument (Cronbach's alpha = 0.85).

**Academic aspirations.** Students' academic aspirations were measured through asking the students what type of education they planned to pursue, where they selected from three response categories (i.e., a. College or university for 3 years or more; b. Short education [1–2 years] after upper secondary school; c. Upper secondary school).

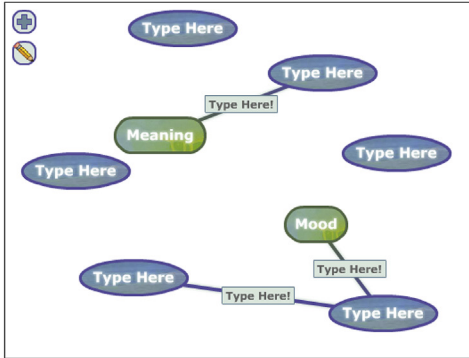
<sup>4</sup> For more information about CoSketch, a multi-user online tool, visit: <http://cosketch.com/>.

**ATCS** ASSESSMENT & TEACHING OF 21ST CENTURY SKILLS

1. GLOBAL HUMAN LEGACY TASK 2011

## My Poem Graphic Organizer

Can you think of some ideas about this poem's **Mood** and **Meaning**?  
Type into **BLUE BUBBLES**, and connect with the **PENCIL TOOL**.



Your Pasted Poem:

Paste Poem Text Here.

VIDEO COLLECTION POEM TEXT TERMS AUTHORS DICTIONARY BASICS

Back Task id: task209 Next

1 2 3 4 5 6 7 8 9

Fig. 7. Sample screen from the mind-map task.

**ATCS** ASSESSMENT & TEACHING OF 21ST CENTURY SKILLS

1. GLOBAL HUMAN LEGACY TASK 2014 NO

## My Poem CoSketch



Log in to CoSketch. Your task is to create a drawing together with your group which reflect the groups' interpretation of the poem. Use the chat tool in CoSketch to discuss the poem with the others in the group.

**Group 62**

After login, choose 'change nickname' in the chat tool and write your user name (e.g., atc001).

Back Task id: Task141081703068 Next

1 2 3 4 5 6 7 8 9

Fig. 8. Sample task from the Human Legacy scenario.

**SES.** Students' socio-economic background was measured by asking the students to indicate the highest level of education their mother or father had attained by selecting between four response categories (i.e., a. College or university degree [3 years



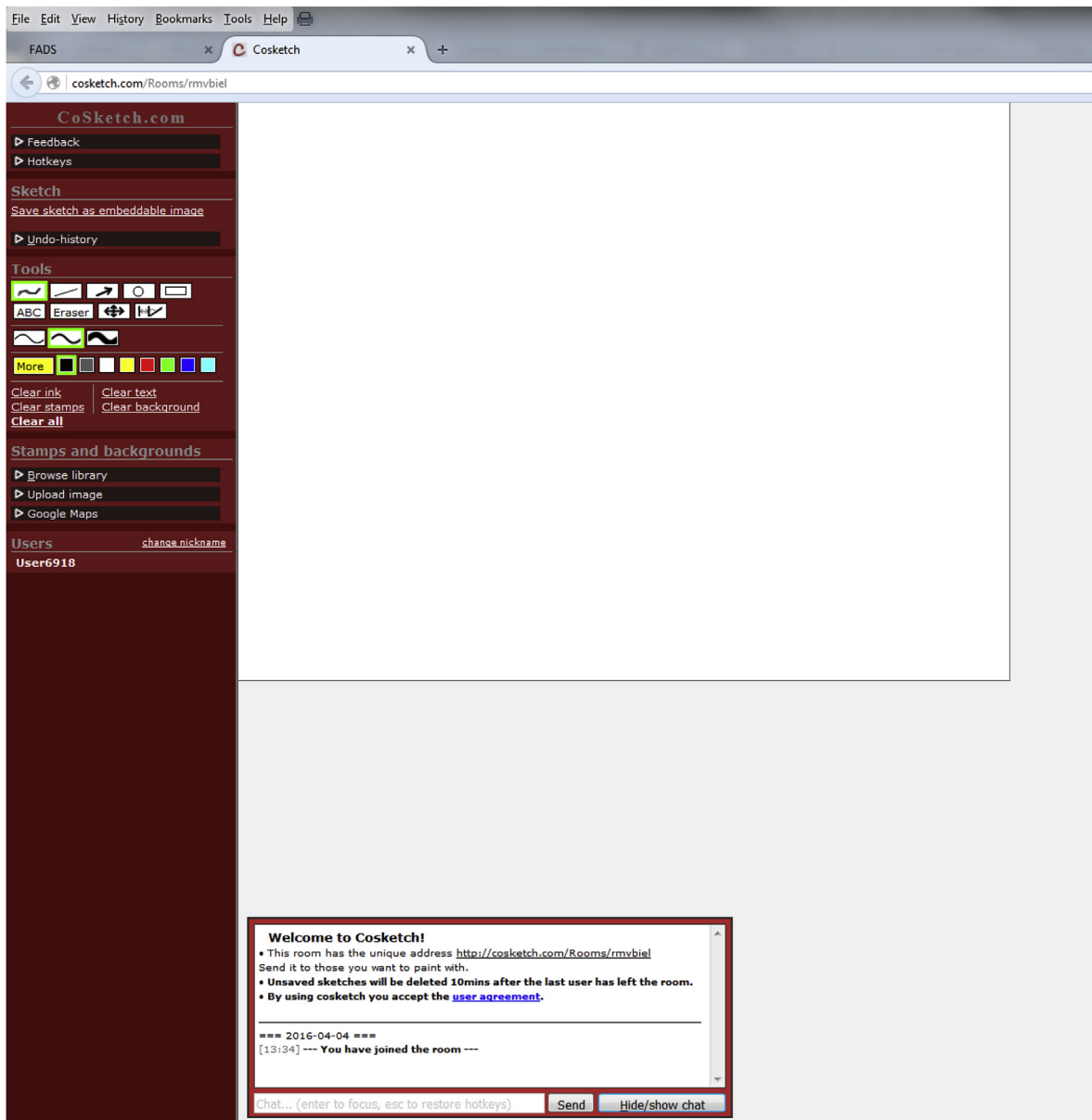


Fig. 9. CoSketch, online multi-user program with drawing and chat functionality embedded.

or more]; b. Short education [1–2 years] after upper secondary [i.e., high school] school; c. Upper secondary school; d. Lower secondary school). For simplicity, only the father's educational level was used when conducting the analyses. However, for 70% of the students, the mother's educational level was the same as the father's.

### 3.4. Sample and procedure

The data were gathered between May and December 2015. A sample of 175 Norwegian students in Grade 9 took the test, and due to missing data, we analyzed a sample of 144 students, of whom 50% were female. With respect to SES (father's education level), 85 students (59%) had fathers in the "college or university for 3 years or more" category, while the remaining 79 students' fathers had either attained education at the level of lower secondary school, upper secondary school (i.e., high school), or short (1–2 years) education after upper secondary school.

The teachers were contacted by email and volunteered their class (es) to participate. The test consists of two scenarios, and the students were allowed 45 min to attempt each. The test was administered by the first author. The software *Conquest* (Adams, Wu, & Wilson, 2012) was used to analyze the data.



### 3.5. Statistical analysis

IRT models are built on the basic idea of latent trait analysis, meaning that they are commonly used to measure individuals' traits that are not directly observable (Baker & Kim, 2004). A number of items may constitute a specific trait (e.g., competences of searching for relevant information, evaluating the information, communication and interaction through digital networks or solving problems collaboratively). Moreover, IRT methodology has been emphasized as a successor to the mere traditional test statistical methods, and the IRT approach has been proven to be specifically valuable for the development and validation of new performance-based tests (Aesaert et al., 2014). Hence, the Rasch model was applied to investigate the reliability and validity of the Norwegian LDN-ICT test and the background questionnaire.

### 3.6. Unidimensional Rasch model

The unidimensional Rasch model (Rasch, 1960) is the simplest IRT model and is based on the following assumptions: sufficiency, monotonicity, unidimensionality, and local independence. See, for instance, Wilson and Gochyyev (2013) for an explanation of these assumptions. In Fig. 10, the structure of the unidimensional model is displayed, consisting of a number of items that load to one composite factor.

The estimation of the models (see Appendix D for details) was done using the marginal maximum likelihood (ML) estimation method, the most commonly used method in IRT. Marginal ML assumes that person-specific parameters are random variables with a particular distribution (Thissen, 1982). Due to the existence of the polytomous items, we used a partial credit model (PCM; Masters, 1982; briefly presented in Appendix D).

### 3.7. Multidimensional random coefficients multinomial logit model (MRCML)

The results obtained from the unidimensional IRT models are valid only to the extent that the unidimensionality assumption is realistic, and the dimensions are distinct. Hence, a set of items measuring two or more distinct latent variables should be analyzed using multidimensional IRT models.

Multidimensional IRT (MIRT) models are categorical variants of confirmatory factor analysis (CFA; McKinley & Reckase, 1983a; Reckase, 1985). MIRT models are considered “full information” factor analysis (Bock et al., 1988) due to the fact that they are fitted to the item responses directly (rather than to polychoric correlations, as in classical factor analysis). MIRT methods have proven to be quite useful in many practical situations and have attracted significant interest.

The multidimensional random coefficient multinomial logit (MRCML) model (a.k.a. multidimensional Rasch model) was proposed by Adams et al. (1997) to analyze data in the Rasch modeling framework (see Appendix D for detailed model descriptions of the MRCML model). The software ConQuest (Adams et al., 2012), which was used in the analysis, implements the MRCML as its core structure. See Adams et al. (1997) for a more detailed presentation of the model and Briggs and Wilson (2003) for an explication of the model.

By using the multidimensional Rasch model, item difficulties, dimension-specific variances, and the covariance between dimensions are obtained. In addition, person-specific predictions of the latent variable locations are obtained. Fig. 11 shows the factor structure of the four-dimensional Rasch model, in which all item loadings for each dimension are constrained to unity (restriction of the Rasch model), and we denote correlations between dimensions with  $\zeta$ .

Note that items can be specified to load on any number of dimensions as long as conditions for the identification of the model detailed in Volodin and Adams (1995) are satisfied.

### 3.8. Differential item functioning

To gather evidence for the internal validity of the instrument across a wide range of respondent variables, such as gender and SES, we investigated whether items function similarly for various categories of these variables. Furthermore, this allows

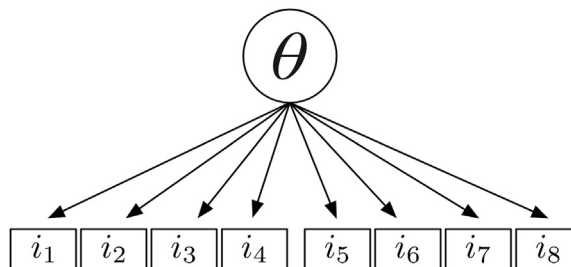


Fig. 10. Unidimensional Rasch model.

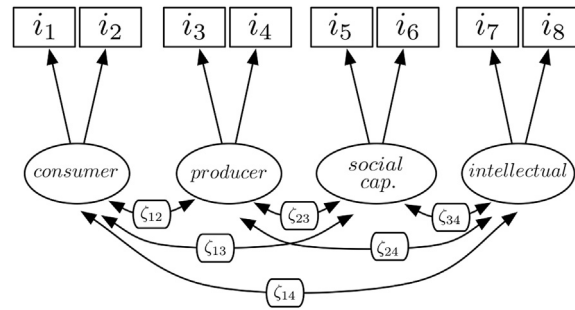


Fig. 11. Four-dimensional Rasch model (for reduced case of two items per dimension).

researchers to study the construct with respect to its external validity by investigating the relations to other constructs and the generalizability of the measurement model across groups.

One of the fundamental assumptions in traditional IRT models is the assumption of local independence. This implies that the respondent's response to a particular item is independent from the responses to other items conditional on the respondent's location on the latent continuum. The consequence of this assumption is that after conditioning the respondent's location on the continuum, the joint probability of item responses is the product of individual item response probabilities (Wilson & Gochyyev, 2013). However, this assumption will not be justified if person-level covariates (such as the respondent's gender or their level of schooling) have direct effects on item responses. The violation of this assumption in such a way signals that the individual item is biased against a particular group(s) and favors another/others and is known as differential item functioning (DIF).

According to Camilli, 2006, (p. 226), DIF is

... said to occur when examinees from groups R and F have the same degree of proficiency in a certain domain, but different rates of success on an item. The DIF may be related to group differences in knowledge of or experience with some other topic beside the one of interest.

One approach to investigating DIF was introduced by Muthén (1985), and it is often used in structural equation modeling. We used an alternative approach, more commonly used in IRT, which relies on the methods introduced by Hambleton and Swaminathan (1985) and Hambleton, Swaminathan, and Rogers (1991). These approaches are based on comparing the relative difficulty of the individual items for each group after accounting for the differences between groups in the overall test (for more technical details on DIF, see Appendix D). In particular, "... an item shows DIF if individuals from different subgroups who have the same ability but they do not have the same probability of getting the item right" (Hambleton et al., 1991, p. 110). For the effect sizes of the DIF statistics, a standard effect size recommended by Longford, Holland, and Thayer (1993) and translated by Paek (2002) into the Rasch model context was considered. According to the recommended rule, a statistically significant logit difference value less than 0.426 is "negligible," a value between 0.426 and 0.638 is considered "intermediate," and a value over 0.638 is considered a "large" DIF.

### 3.9. A note regarding the small sample size

Because of its concise parameterization for item response behavior, the Rasch model requires relatively small sample sizes to obtain accurate parameter estimates. Sample size of about  $N \leq 200$  examinees (Lai, Teresi, & Gershon, 2005; Wright & Stone, 1979) is sufficient to obtain accurate parameter estimates. According to Linacre (1994), the minimum sample size for the item calibration using the Rasch model ranges from 108 to 243 depending on the targeting, with  $N = 150$  sufficient for most purposes.

Using simulations, Paek and Wilson (2011) found that, when there is a DIF, the likelihood ratio test in the Rasch DIF model approach showed higher DIF detection rates than the alternative approach—Mantel-Haenszel chi-square test—for sample sizes of 100–300 per group and test lengths ranging from 4 to 39. Scott et al. (2009) found that detecting moderate uniform DIF in a two-item scale required a sample size of 300 per group for adequate (>80%) power. For longer scales, as is the case with this instrument, a sample size of 200 was adequate.

Lin, Chen, Wu, Yu, and Ouyang (2012), used sample of  $N = 127$  in their study, in which they employed a multidimensional Rasch model (along with DIF detection methods) to validate the dimensionality and reliability of the Frenchay Activities Index—15-item instrument intended to measure two dimensions. With somewhat similar purpose, to investigate the dimensionality and the differential item functioning, Ma, Green, and Cox (2010), used  $N = 177$  using multidimensional Rasch model on a 17-item instrument intended to measure three dimensions.

## 4. Results

### 4.1. Item and dimensionality analysis (RQ1 & RQ2)

In order to address our first research question (e.g., provide evidence for the internal validity), we investigated the unidimensional model consisting of 39 items. For a description of the items and which strands they were assigned to, see [Appendix A](#). Our findings showed a high reliability of 0.86 for the composite across the four dimensions, estimated using the EAP formulation ([Wu, Adams, Wilson, & Haldane, 2007](#)). The EAP estimate is a prediction of the respondent's location in the construct, measured based on his/her responses to the relevant set of items.

We further investigated the multidimensionality of the test. Note that the unidimensional Rasch model is nested in (i.e., a special case of) the multidimensional Rasch model. The difference in deviances obtained from the estimation of the two models is assumed to have a chi-square distribution, with the difference in the number of parameters as degrees of freedom. As a result, we can statistically test whether the less restricted model (multidimensional Rasch model) fits the data significantly better than the simpler model (unidimensional Rasch model). This likelihood ratio test can only be used when the models are nested. Note, however, that the dimension-specific variances cannot be nonnegative; thus the null hypothesis is on the boundary of the parameter space, and thus the LR statistic does not have a simple chi-square distribution. The conservative test can then be obtained by simply dividing the naïve p-value from the LR test by 2 ([Rabe-Hesketh & Skrondal, 2005](#)).

As we see in [Table 3](#) below, the difference in deviances between the two models is 28.8 (4448.1–4419.3) with 9 (57–48) degrees of freedom, which is statistically significant at 0.01.

Thus, we conclude that the four-dimensional Rasch model fits the data statistically significantly better than the simpler unidimensional Rasch model (see [Table 2](#)).

To check whether the items are well aligned with the multidimensional Rasch model, the weighted mean square fit statistic is estimated for each item (also known as item fit). This fit statistic is a measure of the discrepancy between the observed item characteristic curve and the theoretical item characteristic curve ([Wu & Adams, 2013](#)). Ideally, these values are expected to be close to unity, and common conventions of 3/4 (0.75) and 4/3 (1.33) are used as acceptable lower and upper bounds, respectively ([Adams & Khoo, 1996](#)). Values less than one imply that the observed variance is less than the expected variance for that item, while values more than one imply that the observed variance is more than the expected variance for the item. We found that only one of the 39 items in the full version of the instrument (along with step parameters) fell outside the range of 0.75 and 1.33 (fit was 1.34). The item analysis results obtained from the multidimensional Rasch model are summarized in [Table 3](#).

From the set of 39 items, 15 items measure the Consumer dimension, 15 items measure the Producer dimension, four items measure the Social Capital dimension, and five items measure the Intellectual Capital dimension ([Appendix A](#)).

[Table 4](#) shows the estimated latent correlations between the four dimensions obtained from the four-dimensional Rasch model.

[Fig. 12](#), shown below, also known as the Wright map, provides a visual representation of the relationship between the four abilities of respondents and estimated item difficulties by visualizing them on the same scale. The set of “X” represents the distribution of respondents on the relevant dimensions (respondents on the top are estimated to be “higher” on the dimension). The items are represented on the right, with the most difficult at the top and the least difficult at the bottom. When the respondent (“X”) and the item are at the same level, the respondent has approximately a 50% probability of answering the item correctly. When the respondent's location is higher than the location of the particular item, the probability of being successful on that item is higher than 0.5 and vice versa (see [Wilson & Gochyyev, 2013](#) for a detailed explanation).

### 4.2. Differential item functioning (RQ3)

Using the initial set of 39 items, we checked for DIF with respect to gender and SES (proxy for respondent's father's educational level). For confirmation, the test for DIF was conducted using both unidimensional and multidimensional models. Our results showed that two items (from consumer and producer dimensions; items 1 and 9 in [Appendix A](#)) are biased against males, one item is biased against females (item 35), and two items (items 26 and 31) are biased against students whose parents are in the higher education category (i.e., college or university). The differences in difficulty for these items among the groups (gender or SES) were statistically significant, and the effect sizes were large. We eliminated the five items, and in the

**Table 2**  
Model summaries for the unidimensional and multidimensional Rasch model for 39 items.

Model	Deviance	Number of parameters
unidimensional PCM	4869.4	53
4-dim PCM	4840.9	62

**Table 3**

Item analysis results for the multidimensional model with 39 items and 34 items.

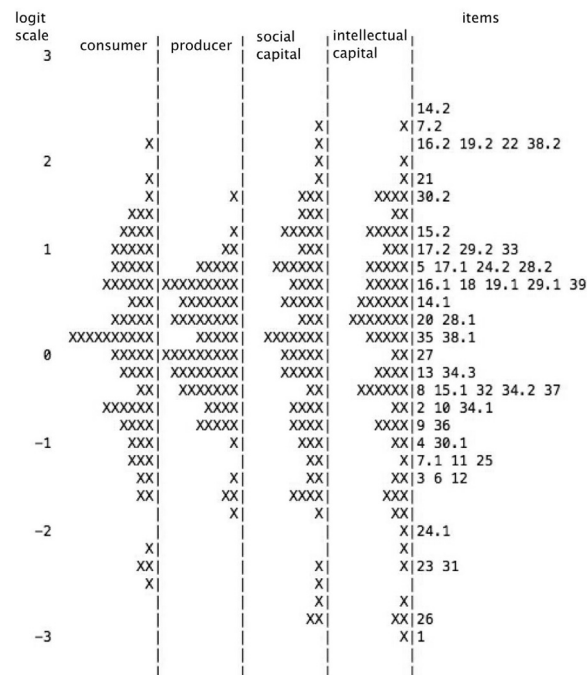
	Multidimensional model (39 items)	Multidimensional model (34 items)
Sample size	144	144
Number of items in calibration	39	34
Number of polytomous items	12	12
Missing data (at random)	30.3%	30.2%
Model	PCM	PCM
Weighted fit MNSQ > 1.33 T sig. (item parms)	1 (1.34)	none
Weighted fit MNSQ > 1.33 T sig. (step parms)	none	none
<i>Reliability estimates:</i>		
EAP/PV reliability		
Consumer (CiN)	0.79	0.84
Producer (PiN)	0.78	0.81
Social Capital (SCN)	0.77	0.71
Intellectual Capital (ICN)	0.70	0.74

Note: The high percentage of missing appeared because not all students responded to the background questionnaire or both scenarios (i.e., Arctic Trek and Human Legacy).

**Table 4**

Correlations among the four dimensions (39 items).

Dimension	Consumer	Producer	Social
Consumer			
Producer	0.90		
Social	0.92	0.91	
Intellectual	0.92	0.76	0.87

**Fig. 12.** Wright map for the four dimensions of the LDN-ICT test. Note: "X" represents the distribution of respondents on the four dimensions.

next set of calibrations we confirmed that none of the differences in the item difficulties between the groups were statistically significant at the 0.05 level for the remaining 34 items.

After removing the five items, we estimated the EAP reliability for the unidimensional test consisting of 34 items, which still showed a high reliability of 0.85. Moreover, we compared the unidimensional and multidimensional models on the reduced scale and established evidence for the four-dimensional model, which also fits the data statistically significantly better than the unidimensional model, as shown in Table 5.

**Table 5**

Model summaries for the unidimensional and multidimensional Rasch model for 34 items.

Model	Deviance	Number of parameters
unidimensional PCM	4448.1	48
4-dim PCM	4419.3	57

Furthermore, the item fit analysis results obtained from the multidimensional Rasch model of the 34-item scale are summarized in Table 3, showing that none of the items fell outside this acceptable range. The estimated correlations among the four dimensions in the multidimensional model varied between 0.74 and 0.92 (Table 6). Moreover, the correlations slightly decreased compared to the correlations in the 39-item multidimensional model (Table 4) except for the relation between Producer and Consumer.

We found that there is not a statistically significant difference in overall performance between males and females. However, students whose father's education level was higher performed statistically significantly higher (at 0.01 level) than their peers with lower paternal education.

#### 4.3. External validity evidence (e.g., relations between test score and other variables; RQ4)

Below we report the results from the unidimensional analysis of the three constructs (ICT Self-efficacy, Collective Efficacy, and Perceived Usefulness of ICT). As shown in Table 7, all three constructs showed high reliability. Further, we present the correlations of the four dimensions obtained from the multidimensional Rasch model with estimated latent variables from the analysis of the questionnaire constructs.

The results of the correlations between the four dimensions and the three constructs measured in the questionnaire are shown in Table 8, indicating varying correlations, ranging between 0.00 and 0.44. The highest correlations were found between students' scores on the Consumer and Producer dimensions and ICT Self-efficacy.

Using Spearman's rho, we checked the correlation of the EAP scores of the four dimensions with the students' responses on a selection of items from each main construct regarding their agreement on different statements. The results are shown in Table 9. The rank-order correlation between item Q21A (agreement on "Being able to learn how to use digital tools will help

**Table 6**

Correlations among the four dimensions (34 items).

Dimension	Consumer	Producer	Social
Consumer			
Producer	0.92		
Social	0.83	0.80	
Intellectual	0.90	0.74	0.85

**Table 7**

Item analysis results for the unidimensional constructs of ICT self-efficacy, collective efficacy, and perceived usefulness.

	ICT self-efficacy	Collective efficacy	Perceived usefulness
Sample size	144	144	144
Number of items in calibration	8	7	6
Number of polytomous items	8	7	6
Missing data	2.6%	1.2%	0.8%
Model	PCM	PCM	PCM
Weighted fit MNSQ >1.35, T sig. (item parms)	none	none	none
Weighted fit MNSQ >1.35, T sig. (step parms)	1 out of 47 (1.39)	none	none
<i>Reliability estimates:</i>			
EAP/PV reliability	0.75	0.82	0.80
Cronbach's alpha	0.79	0.81	0.85

**Table 8**

Correlations between the four dimensions and ICT self-efficacy, perceived usefulness, and collective efficacy (see Appendix B for item descriptions).

	ICT Self-efficacy	Perceived usefulness	Collective efficacy
Consumer	0.29***	0.01	0.06
Producer	0.44****	0.11	0.13
Social	0.28*	0.04	0.04
Intellectual	0.18	0.00	0.00

Note: \*0.1; \*\*0.05; \*\*\*0.01; \*\*\*\*0.001.



**Table 9**

Correlations between the four dimensions and single items from the self-efficacy (Q20A), perceived usefulness (Q21A), and collective efficacy (Q22G) scales and academic aspirations.

Dimensions	Q20A	Q21A	Q22G	Academic aspirations
Consumer	0.26***	0.18**	0.25***	0.18**
Producer	0.28****	0.21**	0.29****	0.21**
Social	0.23***	0.21**	0.22**	0.17**
Intellectual	0.23***	0.19**	0.22**	0.14*

Note: \*0.1; \*\*0.05; \*\*\*0.01; \*\*\*\*0.001.

me in everyday life”) from the Perceived Usefulness construct and EAP scores from the four dimensions showed positive significant relations. Similar results were obtained for items Q22G (Collective Efficacy; agreement on “I’m sure others have a lot to learn from me in terms of technical knowledge”) and Q20A (ICT Self-efficacy; agreement on “I am sure I know how to collaborate with other students by use of digital technology”).

We also checked the correlation of the EAP scores for the four dimensions with the trichotomous variable on *academic aspirations* (Academic Aspirations: “What type of education do you plan to attain?”). As shown in Table 9, we found that the Spearman’s rho measuring the relationship with the ordinal variable is significant at 0.05 for the Consumer, Producer, and Social Capital dimensions and significant at the 0.1 level for the Intellectual Capital dimension.

## 5. Discussion

### 5.1. Internal validity and dimensionality of the test (RQ1 & RQ2)

Our findings regarding the internal validity of the LDN-ICT test showed that none of the 39 items fell outside the acceptable item fit. Moreover, we obtained evidence for high reliability of the full set of items (EAP/PV reliability = 0.86). These results are in line with the reliability indices reported by Wilson and Scalise (2015) for the initial version of the LDN-ICT test. Note that they reported the EAP reliability for the two scenarios separately; Arctic Trek = 0.88 and Human Legacy = 0.93 using EAP. Even though the main analyses in this study were conducted on the overall test, we estimated the EAP values for the two scenarios (Arctic Trek = 0.95 and Human Legacy = 0.92; see Appendix C for results from the unidimensional analysis of the Human Legacy and Arctic Trek instruments as separate tests). Thus, our results consolidate the reliability of the LDN-ICT test.

Furthermore, addressing RQ2, our results showed that the multidimensional model consisting of the four strands (i.e., CiN, PiN, SCN, ICN) fits better than the unidimensional model, and each dimension showed high reliability. These findings lend support to the internal validity of the measure and add evidence of the robustness of the factorial structure across ICT literacy frameworks (Aesaert et al., 2014; Claro et al., 2012; Ferrari, 2013; Griffin et al., 2012). Moreover, these findings are in line with previous research that identified similar facets related to students’ ICT competences. Models that could distinguish between dimensions such as search and retrieval of information, evaluation of information, content creation, and communication showed better fit (Huggins et al., 2014) and were therefore argued as more appropriate conceptualizations of the theoretical framework. Interestingly, similar patterns were obtained in research on the extent to which teachers emphasize developing their students’ ICT competences (Siddiq, Scherer, et al., 2016). Siddiq and colleagues’ results showed that the multidimensional model consisting of the three aspects of ICT literacy—*Accessing*, *Evaluating*, and *Sharing and communicating digital information*—fit better than the unidimensional model. Hence, our findings add evidence to the research literature on the persistence of the structure in the conceptual frameworks of 21st century skills and also the benefits of the incorporation of these distinctions in empirical research. Moreover, the differentiated view on LDN-ICT may therefore reveal the strengths and weaknesses of students’ competences related to each of the four strands and consequently inform classroom instruction and the development of a pedagogical continuum for planning and assessing the learning of 21st century competences (Voogt, Knezek, Cox, Knezek, & ten Brummelhuis, 2011).

Moreover, the empirical evidence for the multidimensional vs. unidimensional model supports the argument for the internal validity of the test, since the hypothesized factor structure was captured and confirmed in the analysis. Furthermore, our analysis showed high correlations between the four dimensions. This finding seems reasonable to some extent, since the four dimensions are all related to the overall concept of learning in digital networks, and each dimension consists of specific competences or activities that are related. In particular, the high correlations between the strands *Consumer* and *Producer* may be due to the fact that the tasks assigned to each of these dimensions were closely related. Such as the task of finding correct information was assigned to the *Consumer* dimension, while using the information and creating content was assigned to the *Producer* dimension. However, these two tasks are closely related, as the first is a pre-requisite for solving the second. The lower correlations between the strands *Social Capital* and *Intellectual Capital* and the strands *Consumer* and *Producer* indicate that they are conceptually distinct. Surprisingly, we identified high correlations between the dimension *Intellectual Capital* and *Consumer*. We assume this may be due to the limited number of items assigned to the *Intellectual Capital* dimension and suggest adding more items to this strand and further investigating the correlations between the strands.

In conclusion, we have provided evidence for the internal validity of the Norwegian LDN-ICT test. Moreover, we have consolidated the validity evidence of the original test in English and have proved that the test also measures the underlying construct in a different language and school context. However, one of the shortcomings in [Wilson and Scalise's \(2015\)](#) study was that, due to small number of respondents, they could not investigate the four dimensions of the LDN-ICT test and therefore mapped all items into one composite variable. Thus, taking a step further, this study provides additional evidence for the conceptualization of the underlying framework and suggests a multidimensional approach to further analysis of the test.

Note that items measuring the consumer or producer dimensions comprise a mix of items coming from both the Human Legacy and Arctic Trek scenarios. This implies that items coming from the same scenario might be more related to each other than items coming from two different scenarios (even though they are both measuring the Consumer dimension). For future research, the models that account for both dimensionality with respect to the four main facets and dimensionality with respect to the scenarios (Human Legacy and Arctic Trek) will be considered. Such a model would require a specification of the six-dimensional model with within-item multidimensionality ([Wang, Wilson, & Adams, 1997](#)). Alternatively, models resembling the testlet ([Wang, Bradlow, & Wainer, 2002](#)) or bifactor models ([Gibbons & Hedeker, 1992](#)) can be considered for such structure. Such models require a larger sample size and thus were not pursued in this study.

### 5.2. Generalizability and differences across gender and SES in LDN-ICT (RQ3)

The DIF analysis across students' gender and socio-economic background revealed DIF in five items (items 1, 9, 26, 31, and 35; [Appendix A](#)). Given the limited sample size, this finding might also be the result of a Type 1 error (incorrectly flagged as DIF). We still flagged and deleted these items. Using the remaining 34 items, we recalibrated the test using the same model (multidimensional partial credit model) and found no evidence that the smaller set of items function differently across gender and SES categories. This result lends additional evidence to the test's validity ([Messick, 1995](#)) and points to the generalizability of the construct. However, for future administration, we will still administer the five items and investigate item bias with the larger sample size. Results regarding DIF tend to be unstable with small sample sizes. For the results to be reliable, it has been suggested, depending on the context, that the minimum sample size be around 100 in the smaller group and 500 in the larger group ([Zwick, 2012](#)). [Davey and Wendler \(2001\)](#) suggested at least 300 for the smaller group and 700 for the larger group. Therefore, we suggest keeping these five items flagged for DIF inspection for the purpose of data collection and further investigating the existence of bias.

After establishing measurement invariance, comparisons across groups (i.e., gender and SES) were possible. Our results did not show differences in overall performance between males and females, and they align with findings from previous research ([Claro et al., 2012](#); [Hohlfeld et al., 2013](#)). However, research on gender effects abounds with conflicting findings, and researchers have reported differences across gender ([Aesaert & van Braak, 2015](#); [Baek et al., 2009](#); [Fraillon et al., 2014](#)). Moreover, a more nuanced picture regarding gender differences was shown in a recent study ([Kim et al., 2014](#)), which indicated that female students' ICT literacy level was higher at the average or lower levels, while male students outperform females at the excellent level. Thus, we compared the variance for male ( $M = 0.83$ ,  $SD = 0.14$ ) and female students ( $M = 0.91$ ,  $SD = 0.15$ ), and our analysis showed that these differences were not statistically significant. Hence, we argue that future research should investigate differences across gender related to both levels and dimensions of the ICT literacy framework.

Regarding differences across students' socio-economic background, our results revealed that students with high SES (i.e., father with higher education) scored higher on the test than students with low SES (i.e., father with no or lower education), and this was significant at the 0.01 level. This finding supports previous research that identified SES as a strong predictor of ICT literacy achievement ([ACARA., 2012](#); [Calvani et al., 2012](#); [Senkbeil, Ihme, & Wittwer, 2013](#)). These findings point toward the digital divide as described by [Van Dijk \(2006\)](#)—the inequality related to ICT literacy and use, which reflects the differences in students' knowledge skills and abilities related to their ICT competences. One of the implications of this finding suggests a continuous need for research on ICT literacy practices within the classroom. Moreover, from an equity point of view, there is a need for studies that provide knowledge about how schools and teachers can support and foster the development of students' 21st century skills.

We note that SES can be measured in several ways (e.g., family income, number of books at home, household resides, occupation), and it may be measured as a proxy or a composite measure that incorporate different indicators of SES ([Oakes, 2016](#)). Hence, the SES measure should be chosen carefully if the study is replicated in a different cultural context.

### 5.3. Relations between LDN-ICT and self-efficacy, collective efficacy, perceived usefulness, and academic aspirations (RQ4)

As a step in investigating the external validity ([Messick, 1995](#)) of the LDN-ICT test, we examined the relations between students' scores on the four dimensions of the test and self-efficacy, collective efficacy, perceived usefulness, and academic aspirations.

The correlations between three of the dimensions of LDN-ICT (CiN, PiN, SCN) and self-efficacy were all positive and statistically significant (as shown in [Table 8](#)). This finding indicates that students who perceive themselves as competent in ICT also tend to score higher on the Consumer, Producer, and Social Capital dimensions. These findings are in line with previous research on the relation between students' ICT achievement and their beliefs in their own ICT competence ([ACARA., 2012](#); [Hohlfeld et al., 2013](#)). We note that the correlation between ICT self-efficacy and the dimension *Intellectual Capital* (ICN)

was positive yet not significant. We assume this could be due to the incompatibility between the two measures. Following Bandura (2006) argumentation on the need for differentiated measures of self-efficacy, we recommend that future research supplement the ICT self-efficacy construct with items that to a larger degree encapsulate the ICN strand.

*Collective efficacy* refers to the individuals' beliefs in their group's skills (Bandura, 2006), and *perceived usefulness* is related to the belief that ICT use would enhance one's performance (Davis, 1989). The correlations between all four dimensions of LDN-ICT and *collective efficacy* and *perceived usefulness* were positive yet small and insignificant. These findings may be due to several reasons, such as the fact that the measures themselves were not sufficiently investigated, as they may contain items that relate to a diversity of ICT usages and tasks. This was evident from the correlations between the LDN-ICT score and the two examples of items we extracted from each of the two constructs (items Q21A and Q22G in Table 9), which indicated statistically significant, high, and positive relations. The fact that some items from each of the measures were significantly related but that the entire measures were not related significantly indicates that the measures themselves are not well established yet. Hence, further investigations of the *collective efficacy* and *perceived usefulness* constructs may benefit from detailed inspection of the constructs (e.g., dimensionality analysis and more alignment of items with the concept measured) and refinement of the scales before further comparisons are made.

As expected, we found significant, high, and positive correlations between the four dimensions of LDN-ICT and *academic aspirations*, indicating that students who aim at pursuing higher education (university or college) also have higher ICT literacy than students who do not plan to attain higher education. These findings are in line with previous research (Fraillon et al., 2014; Hatlevik & Gudmundsdottir, 2013) and support Selwyn's (2009) argument that students' socio-economic background seems to be closely associated with which groups will benefit from technology. Moreover, our findings add to the research reflecting that the digital divide is closely related to the social divide among students (Hatlevik & Gudmundsdottir, 2013).

In sum, we conclude that our results add to the evidence for the external validity of the test by showing positive relations between the test scores and the background variables of self-efficacy, SES, and academic aspirations, as expected.

#### 5.4. Limitations and future directions

The present study has a number of limitations that point to future research in the field of the assessment of 21st century skills. First, due to the small sample size, evidence for a fully invariant measure could not be firmly established. Thus, we suggest that further research on the LDN-ICT test include larger data sets. Second, even though we could verify the four dimensions of the test, we did not elaborate on the links between the different levels of the dimensions. Thus, the model on the hypothesized learning progression (Fig. 1 in the *Theoretical Framework* section) needs to be empirically investigated. Third, the original test included data from students aged 11, 13, and 15, while in the present study only data from 15-year-old students were collected. Hence, future studies may include the development of vertically linked assessments across different age groups to facilitate the study of how students' competences progress over time. This would be helpful for the development of instructional strategies and materials across years of schooling. Fourth, the scoring of the items, especially those concerned with student-student collaboration, were hand-scored following predefined rubrics. On the basis of the experience from this second round of data scoring and analysis, we expect further refinement of the test to incorporate more of the hand-scoring procedures in the test source code itself (e.g., in the program code of the test). Finally, a few items were assigned to the dimensions *Social Capital* and *Intellectual Capital*. This may affect the credibility of the test, since all the aspects of the test may not be equally covered. Hence, we suggest adding more items to these dimensions in further investigations of the test.

Furthermore, one avenue for the further research might be to explore the optimal test design in this context. A comprehensive approach to this would require a separate study with a factorial design. This may include matrix sampling of items and random assignment of respondent to different designs. Then, assuming that the instruments in the different conditions are parallel, groups with different task lengths can be compared with respect to time to completion and performance or any other relevant measure.

One limitation related to modeling was the fact that we have only compared the unidimensional model with the four-dimensional model. The loadings of the items (i.e., which of the dimensions each item loads on) were specified a-priori, making it a confirmatory study. These loadings were based on previous studies (see for instance Wilson, Gochyyev, & Scalise, 2017; Wilson, Scalise, & Gochyyev, 2016). Pulling all available data for the sufficient sample size and employing additional methods of testing the dimensionality must be considered in the future.

Another note regarding the modeling is related to choosing the Rasch model over other approaches such as latent class analysis (see Magidson & Vermunt, 2002; LCA). IRT models (and factor analysis models) are *variable-oriented* approaches since the primary focus is on identifying relations between variables (i.e., dimension structure) and it is believed that these relations apply across all people (Bergman & Magnusson, 1997). In latent class analysis, however, the focus is on studying individuals' based on their response patterns and looking for subtypes (i.e., classes, profiles), thus considered *person-oriented* approach (Bergman & Magnusson, 1997). For the purposes of presenting evidences for the reliability and validity of the new instrument, we chose the former approach. A very interesting and novel research, however, would be to identify different typologies of respondents across the four ICT dimensions (along with the number of profiles and sizes of the profiles) using latent class analysis.

Last, the questionnaire need to be further investigated and require a separate administration to gather evidences for the validity. A separate study focusing on the questionnaire would help to eliminate the redundant and construct-irrelevant items from the set and purify it for the wider use in the future.

## 6. Conclusion and implications

The present study provides evidence for the internal and external validity of the adapted and further developed LDN-ICT test, which measures students' learning in digital networks. In particular, evidence for the *a priori* assumptions on the structure of the construct (i.e., the four strands: CiN, PiN, SCN, and ICN) was demonstrated utilizing the multidimensional IRT approach. Consequently, we argue that a multidimensional view on the construct would be more beneficial for identifying students' strengths and weaknesses related to these dimensions, and thus more targeted interventions could be installed. Additionally, our findings further support the internal validity of the measure, as it proves the consistency of the measure in another language and school environment. Moreover, the relations between LDN-ICT and ICT self-efficacy and academic aspirations advocate the argument for external validity. With respect to differences across students' gender and SES, the test showed satisfying levels of invariance given the small sample size, and support the generalizability of the test, thus adding to the validity argument for the test.

While acknowledging that there is a need for further refinement of the measures, this study has shown promising approaches, and we therefore consider LDN-ICT a test that could be further developed and implemented on a larger scale in future research. Moreover, we support the implementation of this novel assessment for measuring students' 21st century skills, as it includes students' handling of digital information and real-time student-student communication, collaboration, and problem solving. Our study aims to contribute to Wilson and Scalise's (2015) aspiration that "... an important contribution is to encourage more conversation on how information-age trends do not stop at the school door" (p. 79) and points to the novelty and relevance of the LDN-ICT test for 21st century education and assessment.

## Acknowledgement

Special thanks to the Peder Sather Center, for the grant which made the collaboration between the University of Oslo and University of California Berkeley possible. We would like to thank the Berkeley Evaluation and Assessment Research (BEAR) Center for providing access, facilities and help with the translation and revisions of the test. We also want to thank Senior Advisor Jostein Ryen Andresen (University of Oslo) for feedback on lingual and content related issues in the process of translation and preparation of the Norwegian version of the LDN-ICT test.

## Appendix A

An overview of the test items, item descriptions, whether they items were dichotomous or polytomous, and the items for which DIF was indicated.

Nr	Item name	Dimension	Item description	Scoring	DIF
1	c_item1	CiN	Provided correct code from notebook	d	X
2	c_item2a	CiN	Practice, access the correct link (to webpage)	d	
3	c_item2b	CiN	Find correct information on the webpage	d	
4	c_item3	CiN	Access correct link	d	
5	c_item5	CiN	Find and evaluate information	d	
6	c_item6	CiN	Provide correct information	d	
7	c_item7	PiN	Explain answer, reasoning	p	
8	c_item8	ICN	Copy and paste communication	d	
9	c_item9	CiN	Access correct link, find information	d	X
10	c_item10	SCN	Evaluation of team performance	d	
11	c_item11	ICN	Improvement or change—based on team responses	d	
12	c_item12	ICN	Reasoning/explain why you changed your answer	d	
13	c_item13	CiN	Access correct link	d	
14	c_item14	PiN	Find and evaluate information	d	
15	c_item15a	PiN	Explore and select the line that best match the data	p	
16	c_item15b	PiN	What is the function of the sliders	p	
17	c_item16	PiN	Make diagram—choose sections, change size & label	p	
18	c_item18	CiN	Access correct link	d	
19	c_item19	PiN	Reasoning, why information is/is not there	d	
20	c_item20	SCN	Team evaluation	d	
21	c_item21	PiN	Upload document (conversation)	d	
22	c_item22	ICN	Ask question	d	
23	h_item1a	CiN	Access poem by following the link	d	
24	h_item6a	PiN	Reflection, evaluation	p	
25	h_item6	ICN	Ask question	d	
26	h_item1b	CiN	Access the YouTube video	d	X
27	h_item2	PiN	Paste the poem text	d	

(continued)

Nr	Item name	Dimension	Item description	Scoring	DIF
28	h_item3	PiN	Create a mind map using specified tool	p	
29	h_item4	PiN	Add labels to connect moods and meaning of poem	p	
30	h_item7	CiN	Ordering cards	p	
31	h_item12	CiN	Access link to group—CoSketch	d	X
32	h_item13	PiN	Save & upload document (image-file, collaboration)	d	
33	h_item13b	SCN	Upload document (chat file)	d	
34	h_item13e	PiN	Sort cards, supported by the poem or not	p	
35	h_item23c	SCN	Explain—how collaboration changed understanding	d	X
36	h_item24	CiN	Find and write name of a poem	d	
37	h_item25	CiN	Paste link to a poem	d	
38	h_item26	PiN	Knows how to make audio file and explained	p	
39	h_item26b	PiN	Made and uploaded an audio file	d	

*Note:* Item names starting with “c” were in the Arctic Trek scenario, and items starting with “h” were in the Human Legacy scenario. The four dimensions are indicated as follows: CiN = Consumer in Networks; PiN = Producer in networks; SCN= Social capital; ICN= Intellectual capital. Under the category *Scoring*, “d” indicates that the item is dichotomous (e.g., scored with 1 or 0). The “p” indicates polytomous scoring. DIF-category indicates those items that showed differential item functioning.

## Appendix B

Items measuring students' perceived usefulness of ICT, self-efficacy, and collective efficacy.

Item wordings	Reliability
<i>To what extent do you agree with following statements regarding your perceptions of ICT? (1 = Strongly disagree, 6 = Strongly agree)</i>	
<b>Perceived usefulness of ICT</b>	0.80
Being able to learn how to use digital tools will help me in everyday life	
I need good computer skills (ICT competence) to learn other school subjects	
I need good computer skills (ICT competence) to get into higher education (e.g., college or university)	
I need good computer skills to get the job I want	
Good computer skills are required to participate in social development	
Being able to collaborate digitally is important for working more efficiently	
<i>To what extent do you agree with following statements regarding your ICT competences? (1 = Strongly disagree, 6 = Strongly agree)</i>	
<b>ICT self-efficacy</b>	0.75
I am sure I know how to collaborate with other students by use of digital technology	
When an assignment/task requires the use of digital tools, I am confident that I will do a great job	
I am sure I have the ICT competences future education demands	
<i>To what extent do you agree with following statements (for the questions related to collaborative tasks, refer to the group you worked with in the test you just finished) (1 = Strongly disagree, 6 = Strongly agree)</i>	
<b>Collective efficacy</b>	0.82
I'm sure it's easier to solve problems (such as those in the test) if I collaborate with others	
I'm certain my team communicated adequately while working with the different collaboration tasks in the test	
I'm sure my group learned quickly to use the new digital collaboration tools	
I'm sure I have much to learn from others when it comes to using digital tools	
I'm sure others have a lot to learn from me when it comes to using digital tools	
I'm sure I have much to learn from others in terms of content knowledge (mathematics, science, social studies, Norwegian)	
I'm sure others have a lot to learn from me in terms of content knowledge (mathematics, science, social studies, Norwegian)	

## Appendix C

Item analysis results for the Arctic Trek and Human Legacy scenarios.

	Arctic Trek	Human Legacy
Sample size	144	144
Number of items in calibration	22	17
Number of polytomous items	6	6
Missing data	0%	0%
Model	PCM	PCM
Weighted fit MNSQ >1.35, T sig. (item parms)	None	1
Weighted fit MNSQ >1.35, T sig. (step parms)	None	none
<i>Reliability estimates:</i>		
EAP values	0.95	0.92
Cronbach's alpha	0.90	0.85



## Appendix D

### Unidimensional Rasch model (Partial credit model, PCM)

In the PCM, the probability of person  $j$  scoring  $k$  on item  $i$ ,  $P_{jik}$ , can be expressed as

$$P_{jik} = \frac{\exp \sum_{l=0}^k (\theta_j - \delta_{il})}{\sum_{h=0}^{M_i} \exp \sum_{l=0}^h (\theta_j - \delta_{il})}, \quad k = 0, 1, \dots, M_i, \quad (1)$$

where  $\theta_j$  and  $\delta_{ik}$  are the location in the latent variable of person  $j$  and the location (i.e., difficulty) of step  $k$  of item  $i$ , respectively;  $M_i + 1$  is the number of (ordered) categories for the item, and we use the following notational conventions for identification:

$$\sum_{k=0}^0 (\theta_j - \delta_{ik}) \equiv 0, \quad (2)$$

and

$$\sum_{k=0}^h (\theta_j - \delta_{ik}) \equiv \sum_{k=1}^h (\theta_j - \delta_{ik}). \quad (3)$$

For binary items, the PCM simplifies to the Rasch model:

$$P_{ji1} = \frac{\exp(\theta_j - \delta_i)}{1 + \exp(\theta_j - \delta_i)}, \quad (4)$$

or

$$\text{logit}(P_{ji1}) = \tau_{ji1} = \theta_j - \delta_i, \quad (5)$$

in which  $\theta_j$  and  $\delta_i$  are the latent variable estimate of person  $j$  and the location estimate of item  $i$ , respectively.

### Multidimensional random coefficient multinomial logit model (MRCML)

MRCML assumes that a set of  $D$  domain-specific dimensions underlies the examinee responses. MRCML uses two matrices, namely the scoring matrix  $\mathbf{B}$ ,<sup>5</sup> which maps the dimension and relevant items, and a design matrix  $\mathbf{A}$ ,<sup>6</sup> which represents relationships between items and item or step parameters.

Let items be indexed as  $i = 1, 2, \dots, I$  and categories as  $k = 0, 1, \dots, K$ , each item having  $K_i + 1$  response categories. Let  $d = 1, 2, \dots, D$  represent dimensions being measured in  $p = 1, 2, \dots, P$  examinees. Latent variables (i.e., dimensions) will be denoted as  $\theta = (\theta_1, \theta_2, \dots, \theta_d, \dots, \theta_D)$ . Let response patterns be indexed as  $r = 1, 2, \dots, R$ . The random variable  $X_{pik}$  can, then, be expressed such that

$$X_{pik} = \begin{cases} 1 & \text{if response of person } p \text{ on item } i \text{ is in category } k, \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Then,  $\mathbf{X}_{pi} = (X_{pi1}, X_{pi2}, \dots, X_{piK})$  is a binary vector over  $K$  categories, and  $\mathbf{X}_p = (\mathbf{X}_{p1}, \mathbf{X}_{p2}, \dots, \mathbf{X}_{pI})$  is a matrix indicating a response vector for person  $p$ .

Then, the MRCML model can be presented as

$$P(X_{pik} = 1; \mathbf{A}, \mathbf{B}, \xi | \theta_p) = \frac{\exp[\mathbf{b}'_{pik} \theta_p + \mathbf{a}'_{ik} \xi]}{\sum_{k=1}^{K_i} \exp[\mathbf{b}'_{pik} \theta_p + \mathbf{a}'_{ik} \xi]}, \quad (7)$$

in which  $\theta_p = (\theta_{p1}, \theta_{p2}, \dots, \theta_{pD})$  is a  $D \times 1$  vector of dimension parameters for person  $p$ ;  $\mathbf{b}'_{pik}$  is a  $1 \times D$  vector of person  $p$  for item  $i$  and category  $k$  mapping to the dimension  $d$ ;  $\xi$  is a  $m \times 1$  vector of item parameters; and  $\mathbf{a}'_{ik}$  is a  $1 \times m$  vector that represents the

<sup>5</sup> A response of person  $p$  in category  $k$  on factors of item  $i$  is scored  $b_{piks}$ , thus  $\mathbf{b}_{pik} = (b_{pik1}, b_{pik2}, \dots, b_{pikD})$ , representing scoring across  $D$  factors, and  $\mathbf{B}_{pi} = (\mathbf{b}_{pi1}, \mathbf{b}_{pi2}, \dots, \mathbf{b}_{piD})$ , representing scoring matrix for item  $i$  and person  $p$ , and  $\mathbf{B}_p = (\mathbf{B}_{p1}, \mathbf{B}_{p2}, \dots, \mathbf{B}_{pI})$ , representing scoring for person  $p$  across  $I$  items.

<sup>6</sup> Items are described by  $\xi = (\xi_1, \xi_2, \dots, \xi_m)$  vector of  $m$  item parameters. Let  $\mathbf{a}_{ik} = (a_{ik1}, a_{ik2}, \dots, a_{ikm})$  indicate design vector that describes the empirical characteristics of the response category  $K$  of item  $i$ ,  $\mathbf{A} = (\mathbf{a}_{11}, \mathbf{a}_{12}, \dots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \dots, \mathbf{a}_{2K_2}, \dots, \mathbf{a}_{IK_n})$  then being the design matrix.

link between items and corresponding item or step difficulties (for polytomous items). Item parameter vector  $\xi$  is considered unknown but fixed. The vector of latent variable parameters  $\theta_p$  is considered random and assumed to have the multivariate normal distribution with a mean of  $\mu$  and a variance-covariance matrix of  $\Sigma$ , both of which are fixed unknown parameters.

### Differential Item Functioning

Differential Item Functioning (DIF), also known as item bias, is a threat to the validity of the test. Let the response to a given item be represented by  $Y$ , the latent variable be represented by  $\theta$ , and the person-level independent variable be represented by  $Z$  (i.e., gender, race). DIF then can be formally defined as

$$P(Y = y|\theta, Z = z) \neq P(Y = y|\theta),$$

which implies that the value of  $Z$  influences the probability of the response conditional on the latent variable ( $\theta$ ), and hence the violation of the conditional independence assumption.

### References

- ACARA. (2012). *National assessment program – ICT literacy years 6 & 10 report 2011*. Sidney: Australian curriculum. Assessment and Reporting Authority. from [http://www.nap.edu.au/verve/\\_resources/nap\\_ictl\\_2011\\_public\\_report\\_final.pdf](http://www.nap.edu.au/verve/_resources/nap_ictl_2011_public_report_final.pdf).
- Adams, R. J., & Khoo, S.-T. (1996). *Quest*. Melbourne, Australia: ACER Press [computer program].
- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1–23.
- Adams, R. J., Wu, M., & Wilson, M. (2012). *ConQuest 3.0*. Hawthorn, Australia: ACER [computer program].
- AERA, APA, NCME, & (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Aesaert, K., & van Braak, J. (2015). Gender and socioeconomic related differences in performance based ICT competences. *Computers & Education*, 84, 8–25.
- Aesaert, K., van Nijlen, D., Vanderlinde, R., & van Braak, J. (2014). Direct measures of digital information processing and communication skills in primary education: Using item response theory for the development and validation of an ICT competence scale. *Computers & Education*, 76, 168–181. <http://dx.doi.org/10.1016/j.compedu.2014.03.013>.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211. [http://dx.doi.org/10.1016/0749-5978\(91\)90020-T](http://dx.doi.org/10.1016/0749-5978(91)90020-T).
- Ananiadou, K., & Claro, M. (2009). *21st century skills and competences for new millennium learners in OECD countries*. Organisation for Economic Cooperation and Development. EDU Working paper no. 41. Retrieved from <http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP%282009%2920&doclanguage=en>.
- Arnseth, H. C., Hatlevik, O., Kløvstad, V., Kristiansen, T., & Ottestad, G. (2007). *ITU monitor 2007: Skolens digitale tilstand 2007*. Oslo: Universitetsforlaget [itu monitor 2007: The digital state of education: 2007].
- de Ayala, R. J. (2013). The IRT tradition and its applications. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 144–169). New York, NY: Oxford University Press.
- Baek, S.-G., Kim, D., Kim, M.-R., Kim, H. S., Yu, Y. L., Park, S.-H., et al. (2009). Assessing student ICT literacy on a national level. In T. Bastiaens, J. Dron, & C. Xin (Eds.), *Proceedings of e-learn: World conference on e-learning in corporate, government, healthcare, and higher education 2009* (pp. 2269–2275). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).
- Baker, F. B., & Kim, S. H. (2004). *Item response theory. Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.
- Balanskat, A., & Gertsch, C. A. (2010). *Digital skills working Group. Review of national curricula and assessing digital competence for students and teachers: Findings from 7 countries*. Brussels: European Schoolnet.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares, & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 307–337). Greenwich, CT: Information Age Publishing.
- Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, 9, 291–319.
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., et al. (2012). Defining 21st century skills. In Griffin, et al. (Eds.), *Assessment and Teaching of 21st century skills*. [http://dx.doi.org/10.1007/978-94-007-2324-5\\_2](http://dx.doi.org/10.1007/978-94-007-2324-5_2) (© Springer Science+Business Media B.V.).
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3), 261–280.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4(1), 87–100.
- 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. <http://dx.doi.org/10.1016/j.compedu.2011.10.004>.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). American Council on Education and Praeger Publishers.
- Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., et al. (2008). Longitudinal Analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology*, 100, 525–534.
- Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, 81, 78–96.
- Claro, M., Cabello, T., San Martín, E., & Nussbaum, M. (2015). Comparing marginal effects of Chilean students' economic, social and cultural status on digital versus reading and mathematics performance. *Computers & Education*, 82, 1–10. <http://dx.doi.org/10.1016/j.compedu.2014.10.018>.
- Claro, M., Preiss, D. D., San Martín, E., Jara, I., Hinostroza, J. E., Valenzuela, S., et al. (2012). Assessment of 21st century ICT skills in Chile: Test design and results from high school level students. *Computers & Education*, 59, 1042–1053.
- Compeau, D. R., & Higgins, C. A. (1995). Application of social cognitive theory to training for computer skills. *Information Systems Research*, 6(2), 118–143. <http://dx.doi.org/10.1287/isre.6.2.118>.
- Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. Retrieved from: <https://scale.stanford.edu/system/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning.pdf> [30.10.2015].
- Davey, T., & Wendler, C. (2001). *DIF best practices in statistical analysis*. April 3 ([ETS internal memorandum]).
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13, 319–340. <http://dx.doi.org/10.2307/249008>.

- Dede, C. (2009). Comparing frameworks for the 21st century skills. In J. A. Bellanca, & R. S. Brandt (Eds.), *21st century skills: Rethinking how students learn*. Bloomington: Solution Tree Press.
- Durndell, A., & Haag, Z. (2002). Computer self-efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. *Computers in Human Behavior*, 18, 521–535.
- Edmunds, R., Thorpe, M., & Conole, Grainne (2010). Student attitudes towards and use of ICT in course study, work and social activity: A technology acceptance model approach. *British Journal of Educational Technology*, 1–14.
- Ferrari, A. (2013). DIGCOMP: A framework for developing and understanding digital competence in Europe. In Y. Punie, & B. N. Brečko (Eds.), *JRC scientific and policy reports*. Luxembourg: Publications Office of the European Union. <http://dx.doi.org/10.2788/52966>.
- 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*. IEA: Springer Open.
- Fraillon, J., Schulz, W., & Ainley, J. (2013). *International computer and information literacy Study: Assessment framework*. Amsterdam: IEA.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423–436.
- Goddard, R. D. (2001). Collective efficacy: A neglected construct in the study of schools and student achievement. *Journal of Educational Psychology*, 93(3), 467–476.
- Gordon, J., Halsz, G., Krawczyk, M., Leney, T., Michel, A., Pepper, D., et al. (2009). *Key competences in Europe. Opening doors for lifelong learners across the school curriculum and teacher education*. Warsaw: Center for Social and Economic Research on behalf of CASE Network. Retrieved from: [http://www.case-research.eu/upload/publikacja\\_plik/27191519\\_CNR\\_87\\_final.pdf](http://www.case-research.eu/upload/publikacja_plik/27191519_CNR_87_final.pdf).
- Greenlees, I. A., Graydon, J. K., & Maynard, I. W. (1999). The impact of collective efficacy beliefs on effort and persistence in a group task. *Journal of Sports Sciences*, 17, 151–158.
- Griffin, P., & Care, E. (2015). The ATC21S method. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approaches* (pp. 3–33). Dordrecht: Springer.
- Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Dordrecht: Springer.
- Gully, S. M., Incalcaterra, K. A., Joshi, A., & Beaubien, J. M. (2002). A meta-analysis of team efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationship. *Journal of Applied Psychology*, 87(5), 819–832.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12, 535–556.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hatlevik, O. E., & Gudmundsdottir, G. (2013). An emerging digital divide in urban school children's information literacy: Challenging equity in the Norwegian school system. *First Monday*, 18(4). <http://dx.doi.org/10.5210/fm.v18i4.4232>.
- Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills, educational assessment in an information age*. [http://dx.doi.org/10.1007/978-94-017-9395-7\\_2](http://dx.doi.org/10.1007/978-94-017-9395-7_2).
- Hodges, L., & Carron, A. V. (1992). Collective efficacy and group performance. *International Journal of Sport Psychology*, 23(1), 48–59.
- 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, 639–663. <http://dx.doi.org/10.1007/s11423-013-9304-7>.
- Huggins, A. C., Ritzhaupt, A. D., & Dawson, K. (2014). Measuring information and communication technology literacy using a performance assessment: Validation of the Student Tool for Technology Literacy (ST2L). *Computers & Education*, 77, 1–12. <http://dx.doi.org/10.1016/j.compedu.2014.04.005>.
- Keengwe, J. (2007). Faculty integration of technology into instruction and students' perceptions of computer technology to improve student learning. *Journal of Information Technology Education*, 6, 169–180.
- 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. <http://dx.doi.org/10.1016/j.compedu.2014.04.009>.
- Kirkwood, A., & Price, L. (2005). Learners and learning in the twenty-first century: What do we know about students' attitudes towards and experiences of information and communication technologies that will help us design courses? *Studies in Higher Education*, 30(3), 257–274.
- Klassen, R. M. (2004). Optimism and realism: A review of self-efficacy from a cross-cultural perspective. *International Journal of Psychology*, 39, 205–230.
- Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation and the Health Professions*, 28(3), 283–294. <http://dx.doi.org/10.1177/0163278705278276>.
- Lent, R. W., Schmidt, L., & Schmidt, L. (2006). Collective efficacy beliefs in student work teams: Relation to self-efficacy, cohesion, and performance. *Journal of Vocational Behavior*, 68, 73–84. <http://dx.doi.org/10.1016/j.jvb.2005.04.001>.
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.
- Lin, K. C., Chen, H. F., Wu, C. Y., Yu, T. Y., & Ouyang, P. (2012). Multidimensional Rasch validation of the Frenchay Activities Index in stroke patients receiving rehabilitation. *Journal of Rehabilitation Medicine*, 44, 58–64.
- 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* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Magidson, J., & Vermunt, J. K. (2002). *A nontechnical introduction to latent class models. statistical innovations white paper no. 1*. Available at: [www.statisticalinnovations.com/articles/articles.html](http://www.statisticalinnovations.com/articles/articles.html).
- Ma, L., Green, K. E., & Cox, E. O. (2010). Stability of the philadelphia geriatric center morale scale: A multidimensional item response model and Rasch analysis. *Journal of Applied Gerontology*, 29, 475–493.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Koller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality*, 74, 403–456.
- Martin, J. H., Montgomery, R. L., & Saphian, D. (2006). Personality, achievement test scores, and high school percentile as predictors of academic performance across four years of coursework. *Journal of Research in Personality*, 40, 424–431.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- MCETYA. (2007). *National assessment program – ICT literacy years 6 & 10, report 2005*. Carlton South: Ministerial council on education, employment, training and youth affairs. Retrieved from: [http://www.nap.edu.au/verve/\\_resources/2005\\_ICTL\\_Public\\_Report\\_file\\_main.pdf](http://www.nap.edu.au/verve/_resources/2005_ICTL_Public_Report_file_main.pdf).
- McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space*. Iowa City, IA: The American College Testing Program (No. Research Report ONR 83–2).
- Meelissen, M. R. M., & Drent, M. (2008). Gender differences in computer attitudes: Does the school matter? *Computers in Human Behavior*, 24(3), 969–985.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.
- Millap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
- Muthén, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10, 121–132.
- Oakes, J. M. (2016). *Measuring Socioeconomic Status. Behavioral and social sciences research*. E-Source. NIH OBSSR Online Textbook. Retrieved from: <https://obssr.od.nih.gov/wp-content/uploads/2016/05/Measuring-Socioeconomic-Status.pdf>.
- P21 (Partnership for 21st Century Skills). (2012). *Learn for the 21st century. A report and mile guide for 21st century skills. Partnership for 21st Century Skills*. Retrieved from: [http://www.p21.org/storage/documents/P21\\_Report.pdf](http://www.p21.org/storage/documents/P21_Report.pdf).
- P21 (Partnership for 21st Century Skills) and AACTE (American Association of Colleges of Teacher Education). (2010). *21st century knowledge and skills in educator preparation. White paper*. Retrieved from: [http://www.p21.org/storage/documents/aacte\\_p21\\_whitepaper2010.pdf](http://www.p21.org/storage/documents/aacte_p21_whitepaper2010.pdf).

- Paek, I. (2002). *Investigations of differential item functioning: Comparisons among approaches, and extension to a multidimensional context*. Berkeley: University of California. Unpublished doctoral dissertation.
- Paek, I., & Wilson, M. (2011). Formulating the Rasch DIF 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, 1023–1046. <http://dx.doi.org/10.1177/0013164411400734>.
- Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and school achievement. In R. Riding, & S. Rayner (Eds.), *Perception* (pp. 239–266). London: Ablex Publishing.
- Pepper, D. (2011). Assessing key competences across the curriculum—and Europe. *European Journal of Education*, 46(3), 335–353.
- Peterson, E., Mitchell, T. R., Thompson, L., & Burr, R. (2000). Collective efficacy and aspects of shared mental models as predictors of performance over time in work groups. *Group Processes and Intergroup Relations*, 3(3), 296–316.
- Quellmalz, E. S. (2009). In F. Scheuermann, & F. Pedró (Eds.), *Assessing new technological literacies*. In *Assessing the effects of ICT in education. Indicators, criteria and benchmarks for international comparisons*. European Commission. Joint Research Centre. <http://dx.doi.org/10.2788/27419>.
- Rabe-Hesketh, S., & Skrondal, A. (2005). *Multilevel and longitudinal modeling using stata*. College Station, TX: Stata Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research. Chicago: MESA Press (Expanded edition 1983. ed.).
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401–412.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130, 261–288.
- Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., DeGraeff, A., Groenvold, M., et al. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, 62, 288–295.
- Selwyn, N. (2009). Challenging educational expectations of the social web: A web 2.0 far? *Nordic Journal of Digital Literacy*, 2, 72–82. <http://www.idunn.no/ts/dk/2009/02/art04>.
- 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. *Journal for Educational Research Online*, 5, 139–161.
- Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & 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*, 18, 58–84. <http://dx.doi.org/10.1016/j.edurev.2016.05.002>.
- Siddiq, F., Scherer, R., & Tondeur, J. (2016). Teachers' emphasis on developing students' digital information and communication skills (TEDDICS): A new construct in 21st century education. *Computers & Education*, 92–93, 1–14. <http://dx.doi.org/10.1016/j.compedu.2015.10.006>.
- Silva, E. (2008). *Measuring 21st century skills*. Washington: Education Sector Reports.
- Stajkovic, A. D., & Lee, D. (2001). A meta-analysis of the relationship between collective efficacy and group performance. *Paper presented at the meeting of the Academy of Management* (Washington, D.C.).
- Stajkovic, A. D., Lee, D., & Nyberg, A. J. (2009). Collective efficacy, group potency and group performance: Meta-analyses of their relationships and test of a mediation model. *Journal of Applied Psychology*, 94(3), 814–828.
- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175–186.
- Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, 18, 291–307.
- Tsai, P. S., Tsai, C. C., & Hwang, G. H. (2010). Elementary school students' attitudes and self-efficacy of using PDAs in a ubiquitous learning context. *Australasian Journal of Educational Technology*, 26(3), 297–308.
- Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, 39(2), 111–133. [http://dx.doi.org/10.1207/s15326985ep3902\\_3](http://dx.doi.org/10.1207/s15326985ep3902_3).
- Van Dijk, J. (2006). Digital divide research, achievements and shortcomings. *Poetics*, 34(4–5), 221–235. <http://dx.doi.org/10.1016/j.poetic.2006.05.004>.
- Vekiri, I. (2010). Socioeconomic differences in elementary students' ICT beliefs and out-of-school experiences. *Computer & Education*, 54(4), 941–950.
- Vekiri, I., & Chronaki, A. (2008). Gender issues in technology use: Perceived social support, computer self-efficacy and value beliefs, and computer use beyond school. *Computers & Education*, 51(3), 1392–1404.
- Venkatesh, V., Rabah, J., Fusaro, M., Couture, A., Varela, W., & Alexander, K. (2012). Perceptions of technology use and course effectiveness in the age of web 2.0 : A large-scale survey of Québec university students and instructors. In T. Bastiaens, & G. Marks (Eds.), *Proceedings of e-learn: World conference on e-learning in corporate, government, healthcare, and higher education 2012* (pp. 1691–1699). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).
- Volodin, N. A., & Adams, R. J. (1995). *Identifying and estimating a d-dimensional item response model*. Eighth international objective measurement workshop. Berkeley: University of California.
- Voogt, J., Knezek, G., Cox, M., Knezek, D., & ten Brummelhuis, A. (2011). Under which conditions does ICT have a positive effect on teaching and learning? A call to action. *Journal of Computer Assisted Learning*, 29, 4–14. <http://dx.doi.org/10.1111/j.1365-2729.2011.00453.x>.
- Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, 44(3), 299–232. <http://dx.doi.org/10.1080/00220272.2012.668938>.
- Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, 26(1), 109–128.
- Wang, W., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Vol 4. Objective measurement: Theory into practice*. Greenwich, CT: Ablex Publishing.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Wilson, M., & Gochyyev, P. (2013). Psychometrics. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 1–53). Rotterdam, the Netherlands: Sense.
- Wilson, M., Gochyyev, P., & Scalise, K. (2017). Modeling data from collaborative Assessments: Learning in digital interactive social networks. in press *Journal of Educational Measurement*.
- Wilson, M., & Scalise, K. (2015). Assessment of learning in digital networks. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills, volume 2-Methods & Approaches* (pp. 57–81). Dordrecht: Springer.
- Wilson, M., Scalise, K., & Gochyyev, P. (2015). Rethinking ICT literacy: From computer skills to social network settings. *Thinking Skills and Creativity, 21st Century Skills: International Advancements and Recent Developments*, 18, 65–80. <http://dx.doi.org/10.1016/j.tsc.2015.05.001>.
- Wilson, M., Scalise, K., & Gochyyev, P. (2016). Assessment of learning in digital interactive social networks: A learning analytics approach. *Online Learning Journal*, 20(2), 2472–5730.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Wu, M. L., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, 14(4), 339–355.
- Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ConQuest: Generalised item response modelling software (Version 2.0)*. Camberwell, Australia: ACER Press.
- Zwick, R. J. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. RR-12–08) (Princeton, NJ).