MEASURING ICT SKILLS: RELATIONSHIP OF THE ITEM DIFFICULTY AND CHARACTERISTICS OF TEST ITEMS

M.T. Kaarakainen

University of Turku (FINLAND)

Abstract

In this paper the test item characteristics and validity of a performance-based test called the ICT skill test were examined. The test is developed to measure the information and communication technology skills of upper comprehensive and upper secondary school students in Finland. The paper offers a brief overview of ICT skills among upper comprehensive school students and examines what kind of effect the item characteristics have on item difficulty. Based on the results, students performed relatively well in items related to social networking and communication, content creation and using productivity software. Instead, the items requiring more technical skills were found to be challenging for majority of the students. By means of item analysis, the ICT skill test was found to be reliable as a whole and its items being appropriately formed and having acceptable discrimination power. The examination of item characteristics exposed that question text length had no impact on item difficulty. The presence of images and formulas was found to increase item difficulty, whereas UI elements, use of simple text-based multiple-choice questions and on the other hand the possibility to interact with the system were found to decrease item difficulty.

Keywords: ICT skills, measurement, item analysis, item characteristics.

1 INTRODUCTION

The global development and widening availability of information and communication technology (ICT), has made digital skills essential for social interaction, civic participation, information retrieval and processing, academic performance, and professional success [1]. Consequently, according to Aesaert and van Braak [2], these skills should be considered as educational outcomes that students nowadays need to achieve. Therefore, ICT skills need to gain a central role in curriculums on every educational level. This stresses the need for valid assessment instruments in order to measure these skills in education.

The basic starting point for assessing skills in an area are the concepts that are measured. In the case of ICT skills, the definitions, views and frameworks are numerous. In most cases, the concepts consist of a domain part (computer, ICT, Internet, etc) together with a knowledge perspective (competence, literacy, skill, etc) [3]. Van Dijk and van Deursen [4] preferred the pervasive term *digital skills* as it captures the entirety of skills needed to use digital media and services in an information society. Van Laar et al. [5] argue that the concepts used in the research in this area are increasingly taking into account knowledge- or content-related skills intending to widen the traditional device- or technology-centric view of digital or ICT competency.

The new national core curriculum for basic education in Finland aims to offer opportunities for pupils to develop their information and communication technology skills within all subjects. Thus, in this paper, the term ICT skills is used to refer the skills needed to use digital media, services, and devices. This paper concentrates on measurement of ICT skills and presents the background and characteristics of an instrument called *the ICT skill test*, which is a tool for assessing ICT skills of Finnish students. This paper offers a brief overview of the ICT skills among upper comprehensive school students in Finland. Actual research goals in this paper are to confirm, by means of item analysis, the adequacy of the test and to examine what kind of effect the different item characteristics have on item difficulty. Both of these goals serve legitimisation and continuous improvement of the presented instrument.

1.1 Conceptualisation and measurement of digital skills

The test instrument used in ICILS study provides students with an authentic computer-based assessment experience. It applies the construct of computer and information literacy which relates to an individual's ability to use technologies to investigate, create and communicate in order to participate effectively at home or school and in the workplace or society. [6] Another tool, called

ISBN: 978-84-697-9480-7

DIGCOMP, is a self-evaluation tool that uses the concept of digital competence and categorises the crucial competences into information, communication, content-creation, safety and problem solving competencies [7]. In turn, the online survey called Internet Skills Scale (ISS) measures Internet skills, which are found to be distinct from computer skills, as use of the Internet requires more than just abilities to use computers. The ISS consists of self-evaluative questions categorised in the following factors: Operational, information navigation, social, creative, and mobile. [8]

Siddiq et al. [9] summarise that the majority of the tests in previous or current use are self-evaluation tools assessing information search, retrieval, or evaluation, and technical skills, but aspects like problem-solving with ICT, digital communication, and online collaboration are not covered to satisfaction. Similarly, in their systematic literature review, van Laar et al. [5] discovered surveys being the most commonly applied type of study followed by theoretical studies and performance tests. As a result of their review study, they proposed a framework of seven core skills (technical usage, information management, communication, collaboration, creativity, critical-thinking, and problem solving) and five contextual skills (ethical awareness, cultural awareness, flexibility, self-direction, and lifelong learning).

This paper presents an instrument, the ICT skill test, for assessing ICT skills of upper comprehensive and upper secondary education students in Finland. Consequently, the ICT skill test is based on the definitions and goals of ICT competencies in the Finnish national core curriculum for basic education. In the Finnish national core curriculum for basic education (grades 1–9, ISCED 1–2), ICT skills are one of the seven *transversal competences* which are integrated into all subjects, rather than as individual subjects. ICT skills are considered to be an essential factor of civic competencies and are seen both as a goal and a tool for learning. In practice, the goal is to offer every student as follows: (1) Understanding of the basic operations and concepts of ICT, (2) knowledge to use ICT in a responsible, safe and ergonomic manner, (3) skills to use ICT as a tool in information management and creative work, and (4) abilities to use ICT in social communication and networking. In the core curriculum, ICT skills are seen as important civic competencies as such, but also as a key component of other essential competencies like another transversal competence area, multiliteracy. [10]

1.2 Item design in online tests

In their review, Siddiq et al. [9] concluded that the majority of the instruments for assessing ICT skills were based on self-evaluation. A popular type of question in online tests and surveys has been a traditional multiple-choice consisting of a stem and limited options. These type of questions in online surveys or tests have been found to yield higher response rates than open-ended questions [11]. There are seven typical formats of multiple-choice (MC) questions: Conventional MC, alternative-choice MC, true-false MC, multiple true-false MC, matching, complex MC, and context-dependent item set. Conventional MC typically consist of a stem and 3–*n* options, whereas the alternative-choice type MC questions consist only of a stem and 2 options. True-false type of MC questions consist of a declarative statement evaluated in terms of statements truthfulness. Instead, the multiple true-false questions entail a stem with 3 to *n* options and each of these options are evaluated in terms of their truthfulness. Matching type of MC questions consist of a set of options and a set of matching stems. Complex MC questions have a stem which is followed by choices which are grouped into sets for test-takers to choose. Context-dependent item sets in their turn consist of a scenario followed by 1–*n* items and responding to each item is contingent on the scenario. [12] MC questions are simple to create and easy to measure, but they are criticized for excessive attention to memorising and neglecting the aspects of the application of knowledge and skills in real-life situations which emphasize critical thinking or problem-solving skills [13].

While MC questions can be described as a close-ended questions with a limited set of alternatives being offered for respondents, open-ended questions allow the respondent to express themselves without being influenced or guided by given options. In web-surveys open-ended questions tend to produce more missing data and inadequate answers than close-ended questions [11], and they can be laborious to evaluate due to the diverse set of answers. In the field of educational assessment, the concept of *constructed-response* (CR) questions is widely used to refer to a question type that allows for more than one way to correctly answer the questions and which requires test-takers to construct or develop their own answers without suggestions or choices. This kind of question format requires test-takers to apply their knowledge and critical thinking abilities to real-world related problem-solving. CR items can be, for example, tasks in which respondents need to write essays or equations, and describe concepts or processes. [14] CR items which cannot be solved in a traditional pencil and paper format have been described as a *performance assessment* tasks, which may include, for example, conversing in a foreign language or conducting research on an assigned topic [15], but also various kinds of tasks to which a solution requires practical operations [14]. Researchers believe that the combination of MC and CR items increases overall measurement accuracy, as these formats complement each other [16].

Siddiq et al. [17] extracted four task/item design categories when examining instruments developed to measure ICT literacy: Standard multiple-choice (MC), interactive multiple-choice (MC-INT), interactive (INT), and authentic (AUTH), for evaluating task or item design of the ICT literacy tests. Standard MC tasks consist of questions with response options. Likewise, MC-INT items consist of predetermined response formats, but the information provided is dynamic in nature (finding the right answer requires interaction with the test environment, but the task contains response options). INT items require the test-taker to interact with the test environment, for example to input the right values to the input fields or select and click the right function icons and observe the outcomes. The fourth category, AUTH, refers to tasks in which the test-taker solves the tasks in fully authentic situation without dealing with a test environment.

1.3 Item characteristics and item difficulty

Item difficulty is an estimate of the skill level needed to pass a test item. It can be seen as a test-takers' experience. The test-takers' experience about test items' difficulty and test-taking pleasure has been found to relate to item features, like different item formats, knowledge- and content-related features, and cognitive demands, such as reading demands with text-based items [18, 19]. The item text length has been found to predict the item difficulty as the participants' reading comprehension affects their ability to solve text-based tasks [20]. However, the item length may not be a straightforward predictor of difficulty; Stiller et al. [19] found that the length of the item stem was linked to decreased item difficulty where as the length of response options was associated to increased item difficulty.

Usually the presence of images or diagrams is considered to ease tasks and increase test-taking pleasure. Martiniello [20] has proposed that the presence of photographs or drawings in text-based tests significantly reduces item difficulty. Mayer and Moreno [21] argue, that multimedia presentations can result in deeper understanding; when words and pictures are presented together, the learner is more likely to build both verbal and visual representations and to make connections between them. On the contrary, in the study of Stiller et al. [19], the presence of visual images was noticed to increase item difficulty.

Based on the results of Stiller et al. [19], along with the length of the item stem, the presence of tables were associated with decreased item difficulty. They considered tables requiring rather straightforward comprehension, and therefore their presence decreases item difficulty. Instead, the presence of formulas requires technical comprehension and therefore the presence of formulas leads to increased item difficulty [20]. Similarly, the presence of specialist terms increases the item difficulty, as it requires prior-knowledge from the item topic; studies have shown that contexts which students are unfamiliar with tend to lead to increased item difficulty [22].

2 METHODOLOGY

2.1 Participants

The data was collected in Finland at the beginning of year 2017 (January–March) as a project financed by the Finnish Prime Minister's Office (funding for Government's analysis, assessment and research activities). Altogether, 5455 9^{th}-graders ages 15 to 17 years were tested, and of that group 47% were boys and 53% were girls. The mean age for the participants was 15.24. The participants were from 149 upper comprehensive level schools (grade 9/9) in 65 municipalities around the country, chosen based on a geographically representative sample of Finnish municipalities, as formed by the Finnish Education Evaluation Centre.

2.2 Measurement

ICT skills were measured using the online test developed in the Research Unit for the Sociology of Education (RUSE) at the University of Turku (Finland). The original ICT skill test was developed in 2013, and it was completely revised during year 2016 when the Finnish national core curriculum for basic education was renewed (2014) and implemented in schools in August 2016 [23, 10]. The test is bilingual because both Finnish and Swedish are official languages in Finland. The software used for testing is a web application, written in PHP and Javascript and using the TinyMVC- and Bootstrap-frameworks. The application is supported by PostgreSQL database software for all data storage needs, and the test content (tasks and the specific questionnaires in each study) is included in the system as easily changeable XML-files. There are four different levels of user roles in the system: Student, Teacher, Organization, and Admin. Each of these roles has a different set of actions and views available to them.

Data protection issues have been carefully dealt with. Among the previously mentioned roles, the student-role has the least amount of available activities; no login was required (no personal data was stored from minors), as the user entered the test with a simple link. At the beginning of the test, data subject's consent for research was requested (the data subject was able to withdraw the consent at any time during the test session). When the test was done and the user closed the test, all the test data was gathered into a separate research database, and deleted from the original database which is used only for data storage while the test session was active.

The test consists of 18 items divided into 6 modules based on item topics (Appendix 1). The goal was to form coherent modules for a comfortable user experience. Test items were implemented in such a way that the user interface and graphics were intended to simulate common ICT applications and hence mirror real-life settings. The tested competency areas (18 items) were chosen based on the renewed Finnish national core curriculum [10]. The last module (requirements for the ICT study programs) was broadly based on the curricula for the ICT field of Finnish vocational institutions and the universities of applied sciences. The participants could achieve 2 points for each item, which could result in a maximum total score of 36.

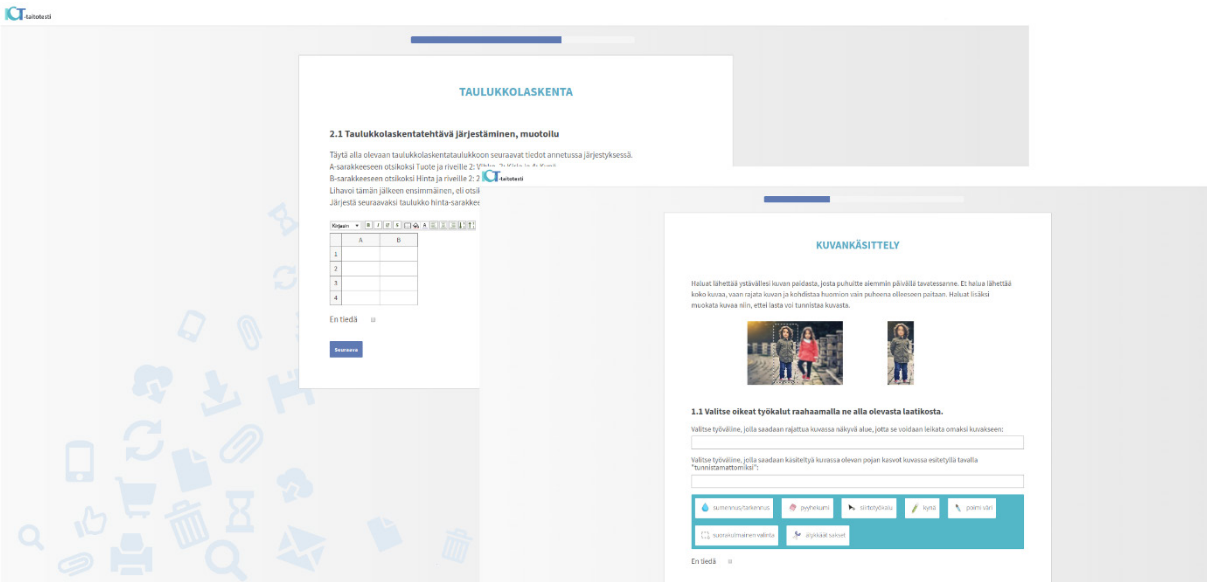

Figure 1. Examples of the two ICT skill test items: Spreadsheets (interactive) and image processing (matching).

In the ICT skill test, each item consists of multiple subtasks (1–6) and/or chains of actions, in which every action (selection or operation) is linked to the previous one; together they form a coherent item. In the items, a combination of close-ended questions (conventional multiple-choice, true-false multiple-choice, multiple true-false multiple-choice, and matching) and open-ended questions or questions requiring participants to interact with the test environment (input the right values or select and click the right function icons) were applied. The majority of items can be seen as context-dependent item sets [cf. 12], as they consist of a problem scenario that participants have to solve by choosing right actions from given options beside the progressive storyline. Some items can be seen as constructed-response questions that require the test-takers to construct or develop their own answers [15]. In the ICT skill test, items are assessed automatically based on specified options and actions or simple text mining algorithms.

2.3 Analysis

The reliability (internal consistency) of the ICT skill test was estimated using Cronbach's alpha. The reliability refers to the extent to which a test is a consistent measure of a specific concept. Along with Cronbach's alpha, the corrected item-total correlation is widely used. Item-total correlation is a correlation of each item score to total score, and it indicates if any item fails to correlate with a total score. It is suitable analysis for items other than extremely easy or difficult items. The cutoff of value .3 means that items that have item-total correlation below the threshold value are likely to be extremely easy or difficult, ambiguous, or otherwise that this item is not measuring the same construct being measured by the other items, and thus its relevance for a test should be carefully inspected. [24]

To analyse item difficulty, an *item difficulty index* was used instead of traditional proportion of right answers. Item difficulty index is considered to be more suitable to be used with open-ended and construct-response items [25], as it suits better for complicated items, where the interest is not simply on how many test-takers get the item (completely) right. The formula used to compute the item difficulty index (*P*) was:

Difficulty index, $P = \frac{\sim \mathrm{fX} - \mathrm{nX_{min}}}{n(X_{\max} - X_{\min})}$

where *~fX* is the total number of scores earned by all test-takers on an item, *n* is the number of test-takers, X_{min} is the smallest item score possible, and X_{max} is the highest item score possible.

The item discrimination is another basic measure of the validity of an item. It is defined as the ability of an item to discriminate (or differentiate) between high achievers and low achievers [26]. The formula used to compute the discrimination index (*D*) is:

Discrimination index, $D = P_U - P_L$

where P_U and P_L are the difficulty indexes for the highest performing (*U*) and lowest performing (*L*) groups. There is no agreement upon the percentage that should be used to determine these groups. The optimum percentage is stated to be approximately 27% in early 1900's, but later anything between 10% and 50% is mentioned in the literature. [26, 27] In this study, the threshold of 27% was used to divide upper and lower groups.

Statistical analysis was carried out using the Python programming language (version 3.6.1) utilizing the following packages: Pandas (library for data analysis, version 0.20.1), NumPy (library for scientific computing, version 1.12.1), and StatsModels (package for statistical computation, version 0.8.0).

3 RESULTS

The Cronbach's alpha for the ICT skill test (all 18 items) was .86, which clearly exceeded the threshold of .7 [24]. As Table 1 presents, the alpha values on the module level varied from .41 to .75. It should be noted that Cronbach's alpha is conservative with a small number of items and particularly scales with two or three items tend to exhibit smaller alphas than those with more items [28]. The relevance of module level reliability is however minor, because the reason to divide test items to 6 modules is more about user-friendly implementation of the content within the user interface than the theory based categorisation of test items.

The scores for modules are shown in Table 1 and for individual items in Appendix 1. In general, all students performed well in items related to information seeking, word processing, social networking, communication, and content creation. The hardest module for the students was, as expected, the requirements in ICT study programs, which included the most technical items and required at least some experience in programming and databases. Applications and basic use modules were also found to be quite difficult for participants. The applications module included items related to purchasing (mobile) applications, installing and updating applications and elementary programming. The basic use module included basic operations, information searching and information networks. Among these, only the information searching was properly mastered, as the students' scores remained weak in both items requiring operational skills. Girls outperformed boys in items related to communication and social networking, using productivity software and content creation tools. Boys in their turn were more successful than girls in items which required more technical knowledge such as basic operations, information networks, software purchasing, installing and updating applications, and in items related to programming.

Table 1. *Reliability and descriptive statistics.*

Scale	Cronbach's alpha: items	Cronbach's alpha: α	Altogether: *M (SD)*	Girls: *M (SD)*	Boys: *M (SD)*	The One-way Analysis of Variance (ANOVA): *F*	*df*	*p*
Total scores	18	.86	10.45 (5.33)	10.64 (4.74)	10.24 (5.92)	7.565	1	.006**
Modules:								
Basic use	3	.48	1.92 (.97)	1.76 (.72)	2.10 (1.17)	171.458	1	.000***
Productivity softwares	3	.66	2.09 (1.65)	2.38 (1.58)	1.77 (1.66)	195.592	1	.000***
Social networking and communication	3	.75	2.45 (1.40)	2.56 (1.32)	2.33 (1.47)	37.799	1	.000***
Content creation and publishing	3	.71	2.30 (1.48)	2.37 (1.40)	2.22 (1.56)	14.643	1	.000***
Applications	3	.52	1.43 (1.11)	1.38 (1.02)	1.49 (1.21)	12.135	1	.000***
Requirements in ICT study programs	3	.41	.26 (.54)	.19 (.43)	.35 (.63)	119.815	1	.000***

**$p < 0.01$, ***$p < 0.001$

Table 2. *Item design and results of item analysis.*

Item	Item Design*	Item Difficulty Index, *P*	Item Discrimination Index, *D*	Corrected Item-Total Correlation, *r*
Basic operations	MC-M	.18	.63	.40
Information seeking	MC	.61	.55	.38
Information networks	MC-M	.17	.57	.33
Word processing	INT	.51	.72	.49
Spreadsheets	INT	.26	.79	.50
Presentations	MC-M	.28	.83	.50
Social networking	MC, MC-M, MC-MTF	.39	.75	.59
Communications	INT, MC-M	.43	.85	.65
Information security	MC-MTF, MC-M	.41	.80	.65
Image processing	MC-M, MC-MTF	.33	.67	.61
Video and audio processing	MC-MTF, MC	.42	.95	.64
Cloud services and publishing	MC-MTF, MC, MC-TF	.41	.93	.56
Software purchasing	MC-M, MC	.19	.58	.45
Installation and updates	MC-MTF	.47	.92	.56
Elementary programming	INT	.06	.57	.40
Database operations	MC-M, MC	.05	.51	.20
Web programming	MC-INT	.08	.34	.23
Programming	MC-M	.01	.44	.22

**MC = multiple-choice, MC-M = matching, MC-MTF = multiple true-false, MC-TF = true-false, INT = interactive*

Results of the item analysis are presented in Table 2. The item difficulty indexes varied from .01 to .61. It is clear purely based on the item mean scores among participants, that items in Requirements in ICT study programs module were difficult for the majority of 9th-graders. Also the elementary programming turned out to be very difficult for participants, even though it actually did not require any programming skills, only the ability to follow and give orders based on given instructions. What was in common with these four items was the presence of some kind of formulas. The *P*-values of items which did not include any formulas were situated within the range of ~.2 and ~.6, indicating that none of these items were either too difficult (<.2) or too easy (>.8) for the participants to complete. The easiest item for participants was the information seeking and the most difficult one was the programming.

The item discrimination indexes ranged between .34 and .92, which are all acceptable (>.3). These values indicate that all items in the ICT skill test are able to distinguish between participants who mastered the skills required in particular item and who did not. The item-total correlation varied between .22 and .65. The last three items in the Table 2, which were among the most difficult items based on item difficulty index, had also the lowest item-total correlation values.

Multiple linear regression analysis was conducted to examine the impact of item characteristics on item difficulty index with the item characteristics as the independent variables and item difficulty as the dependent variable. A significant regression equation was found ($F(9,8) = 8.774$ $p = .003$), with an R^2 of .91. The stem text length had no effect on item difficulty index ($B < -.001$), likewise the option text length ($B < .001$). Presence of images was found to slightly increase item difficulty index ($B = -.051$). In turn, presence of UI elements decreased item difficulty index ($B = .112$). Neither of the last two were not significant. The most evident predictors of item difficulty were found to be possibility to interact with the system, the presence of formulas, and the use of traditional MC type items. The presence of formulas significantly increased item difficulty index ($B = -.156$) whereas the possibility to interact with the system ($B = .200$) and the use of simple text-based MC questions ($B = .265$) significantly decreased item difficulty.

4 CONCLUSIONS

ICT skills are an important tool for learning and a part of modern education. Information and communication technologies are expected to transform and expand learning environments and diversify methods of working. Students are already familiarised to act and learn among digital environments and digital resources in their spare time, and this should be able to connect with the content that is in the focus of formal education. [29] The Finnish national core curriculum [10] aims to foster students to develop their information and communication technology skills within in all subjects by the means of more technology being included in instruction and study. For example, programming has been integrated into the new curriculum as part of the objectives set for mathematics (grades 7–9/9) and the fundamentals of programming in the lower grades (1–6/9) as part of all subjects.

The paper discovers that students performed relatively well in items related to social networking and communication, content creation and using productivity software. Instead, the items which required more technical skills, like programming, basic operations or information networks, were found to be challenging for the majority of ninth graders. This result is interesting, because in many previous studies, young people are identified as skillful users, and are thought to be proficient particularly in operational skills [e.g. 4]. This paper reveals that young people in Finland lack basic knowledge and usage skills of computers. This may be due to the increased use of simple-to-use mobile devices mainly for communication and entertainment purposes instead of more complicated devices among youngsters. Girls were found to be more proficient than boys in communication and social networking, content creation and using productivity software. Boys in their turn dominated the items of basic operations, information networks, software purchasing, installing and updating applications, and all items related to programming. Still, even the majority of boys were found to perform inadequately in items requiring operational and technical knowledge.

Based on the result of item analysis, the items of elementary programming, database operations, web programming, and programming should be taken under consideration. Here it should be noted that as the ICT skill test, with the same content, is used for students both in upper comprehensive school final graders (9/9) and in upper secondary education, it understandably means, that at least some test items are difficult for the youngest target group. In addition, the ICT skill test is intended to test skills needed in real-world problem-solving situations, where there exists no differences based on actors' age. However, since August 2016, programming has been integrated into the Finnish national core curriculum. Hence, all items will be kept in the future, as they provide the ability to track whether the desired programming skills will increase in the future or will not. Along with some item level improvements, the most crucial improvement need in future concerns the time needed to pass the test, as at the moment, the test is rather laborious (execution time: *min* 9.02, *max* 50.52, *M* 25.95, *SD* 8.83). To shorten the test, the analysis of which set of the ICT skill test's explanatory variables (items) are enough to acceptably cover the total ICT skills is needed in future.

The examination of item characteristics exposed unexpectedly that question text length, neither the stem or option length, had no impact on item difficulty. The presence of images was found to slightly increase item difficulty, which is in line with Stiller et al. [19]. Even if images are usually expected to ease tasks or to result in deeper understanding [21], it seems that images may also be related to increased item complexity and these tasks require skills of multiliteracy, which can lead to increased difficulty. The presence of formulas clearly increased item difficulty, as they required technical comprehension. This discovery is in line with previous research [20]. Both UI elements and especially the possibility to interact with the system were found to decrease item difficulty. These kinds of items simulate real-life context, and at best, allow test-takers to try and observe the results of actions instead of leaning on memorising. The traditional text-based multiple-choice questions were also found to decrease item difficulty. These types of questions seem to require rather straightforward comprehension, and therefore they are not as challenging for the test-takers.

As Siddiq et al. [18] pointed out, even though the core ICT competencies may be relatively stable over time, the content on these competencies is under ongoing changes due to rapid technological innovations. These changes force assessment instruments to be based on continuous development, when aiming the instrument to meet the demands of current external content and technological milieu at each given time. As a consequence, the assessment instruments need to be in a continuous cycle of development, operation, evaluation, and adjustment. This brings the development of these ICT-related research instruments very close to agile research and development - or even experiment-driven [30] approach of software development.

ACKNOWLEDGEMENTS

This research and the development of the test instrument was supported by the Finnish Prime Minister's Office (project: "Comprehensive school in the digital age") and the Strategic Research Council at the Academy of Finland (project: "Occupational restructuring challenges competencies").

REFERENCES

[1] L. H. Hoffman & A. L. Schechter, "Technical skills required: how technological efficacy influences online political behavior," *Journal of Broadcasting & Electronic Media*, vol. 60, pp. 484–502, 2016.

[2] K. Aesaert & J. van Braak, "Gender and socioeconomic related differences in performance based ICT competences," *Computers & Education,* vol. 84, pp. 8–25, 2015.

[3] O. E. Hatlevik, G. Ottestad & I. Throndsen, "Predictors of digital competence in 7th grade: a multilevel analysis," *Journal of Computer Assisted Learning*, vol. 31, no. 3, pp. 220–231, 2015.

[4] J. A. G. M. van Dijk & A. J. A. M. van Deursen, *Digital skills. Unlocking the information society,* New York: Palgrave MacMillan, 2014.

[5] E. van Laar, A. J. A. M. van Deursen, J.A.G.M. van Dijk & J. de Haan, "The relation between 21st-century skills and digital skills or literacy A systematic literature review," *Computers in human behavior*, vol. 72, pp. 577–588, 2017.

[6] J. Fraillon, W. Schulz & J. Ainley, *International computer and information literacy study: Assessment framework,* Amsterdam: IEA, 2013.

[7] A. Ferrari, A. *DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe.* Luxembourg: European Union, 2013.

[8] A. J. A. M. van Deursen, E. J. Helsper & R. Eynon, "Development and validation of the Internet Skill Scale (IIS)," *Information, Communication & Society*, vol. 19, pp. 804–823, 2016.

[9] F. Siddiq, O. E. Hatlevik, R. V. Olsen, I. Throndsen & R. Scherer, "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*, vol. 19, pp. 58–84, 2016.

[10] FNBE, *National Core curriculum for basic education 2014.* Helsinki: FNBE, 2016.

[11] U. Reja, K. L. Manfreda, V. Hlebec & V. Vehovar, "Open-ended vs. close-ended questions in web questionnaires," *Developments in Applied Statistics (Metodološki zvezki)*, vol. 19, pp. 159–177, 2003.

[12] T. M. Haladyna, S. M., Downing & M. C. Rodriguez, “A Review of multiple-choice item-writing guidelines for classroom assessment,” *Applied Measurement in Education*, vol. 15, no. 3, 309–334, 2002.

[13] T. M. Haladyna, *Developing and validating multiple-choice test items,* 3rd edition, London: Routledg, 2011.

[14] T. B. Miller, *Using Differential item functioning to test for inter-rater reliability in constructed response items,* Theses and Dissertations, Milwaukee: University of Wisconsin-Milwaukee, 2015.

[15] S. A. Livingston, *R & D Connections. Constructed response test questions: Why we use them, how we score them,* No 11, Princeton, NJ: Educational Testing Service, 2009.

[16] W. Wang, F. Drasgow & L. Liu, “Classification accuracy of mixed format tests: a bi-factor item response theory approach,” *Frontiers in Psychology*, vol. 7, 2016.

[17] F. Siddiq, P. Gochyyev & M. Wilson, “Learning in digital networks – ICT literacy: A novel assessment of students’ 21st century skills,” *Computers & Education*, vol. 109, pp. 11–37, 2017.

[18] M. A. Lindner, J. M. Ihme, S. Saß & O. Köller, ”How representational pictures enhance students’ performance and test-taking pleasure in low-stakes assessment,” European Journal of Psychological Assessment, 2016, doi: 10.1027/1015-5759/a000351

[19] J. Stiller, S. Hartmann, S. Mathesius, P. Straube,R. Tiemann, V. Nordmeier, D. Krüger & A. Upmeier zu Belzen, “Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty,” *Assessment & Evaluation in Higher Education*, vol. 41, no. 5, pp. 721–732, 2016.

[20] M. Martiniello, “Linguistic complexity, schematic representations, and differential item functioning for english language learners in math tests,” *Educational Assessment,* vol. 14, pp. 160–179, 2009.

[21] R. E. Mayer & R. Moreno, “Aids to computer-based multimedia learning,” *Learning and Instruction*, vol. 12, 107–119, 2002.

[22] F. Le Hebel, A. Tiberghien & P. Montpied, *Sources of difficulties in Pisa science items. Proceedings of the European Science Education Research Association Conference 2013*. Nicosia: ESERA, 2013.

[23] M.-T. Kaarakainen, O. Kivinen & T. Vainio, “Performance-based testing for ICT skills assessing: a case study of students and teachers’ ICT skills in Finnish schools,” *Universal Access in the Information Society,* 2017*,* doi: 10.1007/s10209-017-0553-9.

[24] J. C. Nunnally & I. H. Bernstein, *Psychometric theory,* 3rd edition, New York: McGraw-Hill, 1994.

[25] D. T. Tiruneh, M. De Cock, A. G. Weldeslassie & R. Janssen, “Measuring critical thinking in physics: development and validation of a critical thinking test in electricity and magnetism,” *International Journal of S*M. Murray-Ward, *Educational measurement. Origins, Theories and Explications,* Volume II, Theories and Applications, Lanham: University Press of America, 1996.

[26] W. K. Adams & C. E. Wiemanc, “Development and validation of instruments to measure learning of expert-like thinking,” *International Journal of Science Education*, vol. 33, no. 9, pp. 1289–1312, 2011.

[27] R. A. Peterson, “A Meta-analysis of Cronbach’s coefficient alpha,” *Journal of Consumer Research*, vol. 21, no. 2, pp. 381–391, 1994.

[28] S. Hänninen-Vahtivuori, I. Halinen, H. Niemi, J. Lavonen & L. A. Lipponen, “New Finnish national core curriculum for basic education (2014) and technology as an integrated tool for learning”, in *Finnish Innovations & Technologies in Schools* (H. Niemi, J. Multisilta, L. Liponen & M. Vivitsou eds.), Rotterdam: Sense publishers, pp. 21–32, 2014.

[29] H. Holmström Olsson, J. Bosch & H. Alahuari, “Towards R&D as innovation experiment systems: A framework for moving beyond agile software development”, in *Proceedings of the IASTED International Conference on Software Engineering*, pp. 798–805, 2013.

APPENDIX 1. ITEMS IN THE ICT SKILLS TEST

Item	Description	Score (0–2) *M (SD)*
Basic operations	Participants have to pair a keyboard shortcut with a correct action and choose a correct type of computer memory for present education situation.	.36 (52)
Information seeking	Participants have four cases where they have to choose a correct source/channel, out of three, on where to further seek information on a given topic. After this, they are presented with list of search engine results and are asked to choose relevant items related to given scenario.	1.21 (.45)
Information networks	Participants are given four network usage scenarios and have to pair them with correct data transmission technologies and then match correct descriptions of computer network-related concepts.	.34 (.41)
Word processing	Participants are asked to edit (bold, italicise, underline and highlight) a given sample text.	1.02 (.82)
Spreadsheets	Participants are asked to fill a spreadsheet table with given information, bold a header row, and sort the table in ascending order.	.51 (.63)
Presentations	Participants are given a general user interface view of presentation software, with essential sections marked. The task is to pair a correct name with the right section of this view.	.56 (.67)
Social networking	Participants have to pair correct social networking services with four service descriptions, define the meaning of social networking service, and choose four items out of nine that relate to the security of social networking services.	.77 (.51)
Communication	Participants have to fill in the receiver fields, carbon copy, and blind carbon copy) of an email and add an attachment according to instructions, and identify the types of information that can be used to identify Internet users.	.86 (.62)
Information security	Participants have to choose correct statements for secure network communications and choose from alternatives those that are related to the information security of computers in an Internet cafe abroad.	.81 (.58)
Image processing	Participants have to select correct image processing tools for cropping an image and make the person appearing in the image unrecognisable. Afterwards, participants have to choose correct image processing using related statements from given options and choosing the correct file formats for vector graphics.	.65 (.46)
Video and audio processing	First, participants have to choose those methods that can be used to edit video footage from a single camera and then choose a right answer to the question: “Which one of these alternatives is related to lossy audio compression?”.	.84 (.68)
Cloud services and publishing	In the first step, participants have to choose which of the given statements about cloud services are true. In the second step one must choose the correct YouTube-video sharing option that enables limited sharing even to those who do not have an account on YouTube. The third step is a continuation question: “Can we now be certain the video does not circulate to the rest of the Internet for outsiders to see [...]?”	.81 (.69)
Software purchasing	Participants have to choose which matters need to be considered when evaluating the information security of mobile applications and also choose the correct definition of personal data protection from four alternatives.	.38 (.43)
Installation and updates	In the first step, participants choose whether a statement is about an installation or an upgrade and in the second step, they choose whether a statement is related to an update or an upgrade.	.93 (.70)
Elementary programming	Participants have to write, per instructions, a maze traversing script that leads from the starting point to the end. After this, they are presented a short pseudo-code and they have to write the value of a particular variable after the given code has completed.	.12 (.37)
Database operations	Participants have to form an SQL-query, based on given instructions and a simple database diagram, then choose the right definition for the concept ‘NoSQL database’.	.10 (.30)
Web programming	Participants are given three files (HTML, CSS and JavaScript) to use to create a website and the view generated by these files. Participants then answer four multiple choice questions to edit the simple web page view and the dependencies between these given files.	.15 (.32)
Programming	The programming task requires the participants to place lines of Java code in the correct places based on given comment sections.	.01 (.13)