Experimental Validation Strategies for Heterogeneous Computer-based Assessment Items

Lena Engelhardt[a], Frank Goldhammer[a,b], Johannes Naumann[c], Andreas Frey[d,e]

[a]German Institute for International Educational Research (DIPF), Frankfurt am Main (Germany)

[b]Centre for International Student Assessment (ZIB), Germany

[c]Goethe University Frankfurt am Main, Germany

[d]Friedrich Schiller University Jena, Germany

[e]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

Correspondence concerning this article should be addressed to:

Lena Engelhardt, German Institute for International Educational Research (DIPF)

Schlossstraße 29, 60486 Frankfurt am Main, Germany

Phone: +49 69 24 708 754

Email: lengelhardt@dipf.de

Abstract

Computer-based assessments open up new possibilities to measure constructs in authentic settings. They are especially promising to measure 21st century skills, as for instance information and communication technologies (ICT) skills. Items tapping such constructs may be diverse regarding design principles and content and thus form a heterogeneous item set. Existing validation approaches, as the construct representation approach by Embretson (1983), however, require homogenous item sets in the sense that a particular task characteristic can be applied to all items. To apply this validation rational also for heterogeneous item sets, two experimental approaches are proposed based on the idea to create variants of items by systematically manipulating task characteristics. The *change*-approach investigates whether the manipulation affects construct-related demands and the *eliminate*-approach whether the test score represents the targeted skill dimension. Both approaches were applied within an empirical study ($N = 983$) using heterogeneous items from an ICT skills test. The results show how changes of ICT-specific task characteristics influenced item difficulty without changing the represented construct. Additionally, eliminating the intended skill dimension led to easier items and changed the construct partly. Overall, the suggested experimental approaches provide a useful validation tool for 21st century skills assessed by heterogeneous items.

Keywords: validation; experimental strategies; heterogeneous item sets; computer-based assessment; ICT skills

Highlights:

- two experimental validation strategies are proposed

- manipulating of task characteristics for validation purpose

- suitable for heterogeneous computer-based assessment items

- combine experimental understanding with recent developments in validity research

Assessments are increasingly carried out by means of computers enabling the automatic evaluation of responses, and more efficient (i.e., adaptive) testing. With the advance of computer-based assessment, there is an ongoing and pertinent debate around the validity of the test score interpretation, as computer skills are required to complete the tasks. This is true even in domains where it would appear naturally at first sight to use the computer as an assessment tool, because the targeted skill unfolds in a digital environment as well, such as digital reading (see OECD, 2011). Even in these domains, extra care needs to be taken that the assessment targets individual differences in reading-related processes, and not merely computer skills. Actually, for most so-called 21st century skills (e.g., problem solving, collaboration, information literacy; Binkley, Erstad, Herman, Raizen, Ripley, & Rumble, 2012) computers are needed in order to measure them in realistic settings and specifically, simulated environments provide more authentic task settings.

Also educational large-scale-assessments such as PISA (Programme for International Student Assessment; OECD, 2014) or PIAAC (Programme for the International Assessment of Adult Competencies; OECD, 2012) are nowadays assessed by means of computers. Computer-based assessments allow assessing skills performance-based, by not only asking for instance "how good are you in digital reading?" but asking students to actually perform digital reading tasks. Besides, test scores from such studies are interpreted more general in terms of requirements for societal participation (e.g., in PISA). They are not in the first place based on conventional psychological constructs such as intelligence, but on "institutionally defined knowledge domains" (Watermann & Klieme, 2002, p. 2). To be able to justify such a far ranging test score interpretation, typically broad constructs and – in turn – heterogeneous items representing a wide range of contents and situations are needed. Items in these educational studies differ more strongly from each other than items used to assess conventional psychological constructs. This is because items are often instructed and

designed in a contextualized way within a certain situation. They differ in their appearance, but also in the demands and in the knowledge they require (for sample questions see e.g., https://www.oecd.org/pisa/test/form/). With „heterogeneous", we refer to a property of items measuring a certain construct, but not to assumptions regarding the underlying dimensional structure. An example for such a broadly defined but one-dimensional 21st century skills construct is computer and information literacy assessed in the computer-based ICILS-Study (International Computer and Information Literacy Study; Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

Such "innovative item formats" have obvious advantages in terms of construct representation (Sireci & Zenisky, 2006, p. 329) because items can be then more contextualized or authentic, but give rise to new challenges for the validation of test score interpretations (e.g., Linn, Baker, & Dunbar, 1991), as further skills, for instance skills to interact with a computer environment, are involved in the task solution process especially in performance-based assessments. Thus, validation needs to take this into account by providing evidence that the assumed construct-related processes are actually exercised by the test-taker. The validity-threatening potential of such skills is even more an issue when domains are being assessed with the computer that by themselves have no overlaps with ICTs, such as print reading, science, or mathematics. In traditional correlational approaches, these validity threats are addressed by including additional measures in the validation design that directly assess computer or ICT skills. Thereby, discriminant evidence can be provided (see AERA, APA, & NCME, 2014)

The goal of this paper is to present two experimental validation strategies that offer an additional way of dealing with the issue of validity in computer-based assessment and can be also applied to heterogeneous item sets. In the following section, we first briefly describe Embretson's (1983) construct representation approach as the two suggested approaches are

based on this validation rationale. We refer then to how the two experimental validation strategies proposed in this article differ from Embretson's approach. They are described in term of their conceptual basis and also in terms of concrete consequences for building hypotheses in the validation process. An application of the two approaches is presented using empirical data gathered with a test measuring information and communication technology (ICT) skills.

1. Embretson's construct representation approach

Validity is not a property of a test but of "the interpretations of test scores for proposed uses" (cf. AERA, APA, & NCME, 2014, p.11). Kane (2013) suggests that especially a theory-based interpretation, for instance related to the construct, requires ambitious claims of validity. Thus, the strategy for validating a test score interpretation depends on the intended use and the inferences that should be made based on the test scores. A very important and also ambitious claim for justifying the construct interpretation would refer to the relation of task characteristics to the test-taker's score based on the underlying process model that is derived from theory (Kane, 2013). Such claims can be investigated using Embretson's construct representation approach.

The rationale behind the construct representation approach is to determine task characteristics that should theoretically evoke the targeted cognitive processes. These task characteristics – that should also have guided the item development process – are then related to task performance, for instance to item difficulty. If items showing those task characteristics to a greater extent are also harder, test scores can be interpreted as determined by the targeted construct. An example for such a task characteristic could be the number of transformations in a mental rotation task that describes the items' complexity (cf. Embretson, 1983) or the number of orthographic neighbors in a word recognition task, thus words that differ in their

spelling from the targeted word in only one letter. Such task characteristics can be described as complexity factors that can be quantified and describe the complexity of an item in terms of cognitive processes that have to be performed. Thus, the approach refers to the cognitive processes that are assumed to occur while working on the task.

The construct representation approach was applied in many studies, for instance for mental rotation tasks (Caissie, Vigneau, & Bors, 2009), problem solving tasks (Greiff, Krkovic, & Nagy, 2014), computer simulated microworlds (Stadler, Niepel, & Greiff, 2016), or reading comprehension tasks (Hartig & Frey, 2012). Note that in these studies, the stimulus material was homogeneous, that is all items could be described by the same stimulus characteristics in the items. In the mental rotation task, for instance, stimuli have to be evaluated whether they represent a rotation of the initial figure or not, which leads to items with comparable stimulus materials and task solution processes. Defining comparable task characteristics across items, however, might be only feasible in more restricted domains that are not as broad as some domains that are assessed in large-scale assessments (Watermann & Klieme, 2002). This holds for instance for the ICT skills test used in this study, because users have to deal with, for instance, different applications (e.g., browser or e-mail) and different information tasks (e.g., access or evaluate information).

Combining one type of information tasks with one environment might in fact compose a facet of ICT skills that can be measured with a homogeneous items set, making it possible to employ Embretson's (1983) construct representation approach. For example, Pfaff and Goldhammer (2011; see also Hahnel, Goldhammer, Naumann, & Kröhne, 2016) described a test measuring the evaluation of information presented in browser environments. In this case, item features can be identified that are comparable across all items, as for instance the number of to-be-accessed hypertext pages. However, when a comprehensive assessment of ICT skills is intended, the different information tasks and the different applications imply that

it will be difficult to find task features that can be defined for all tasks in the assessment alike. Think, for instance, of a task requiring information to be created using computers. Such a task might require changing font sizes or the position of text fields in a presentation. A task requiring information to be accessed in contrast might require the test-taker rather to navigate text presented in a browser environment. The two suggested experimental approaches make the rational of relating task characteristics to item difficulty also feasible for heterogeneous item sets.

Embretson (1983) describes besides "construct representation" a second approach to validation, the "nomothetic span" approach. While "construct representation" focuses on task differences, "nomothetic span" targets individual differences. In the nomothetic span approach, the relations to other constructs as predicted by the nomological network or that are supposed to underlie the item solution process are investigated. The idea of a nomological network is to find evidence for supporting the targeted test score interpretation for a specific use by investigating the relation to other variables. These can be variables that are assumed to be related (convergent evidence) and variables that are assumed not to be related to the test scores (discriminant evidence) (cf. AERA, APA, & NCME, 2014).

2. Two new experimental validation strategies

We want to apply Embretson's approach (1983) of relating task characteristics to task performance also to heterogeneous item sets, where potentially every item belongs to a separate item type. Such item sets are frequently used in computer-based (large-scale) assessments of student achievement to measure broadly defined constructs. The novelty of the two proposed experimental approaches is to systematically construct variants of original items by manipulating certain task characteristics for validation purposes. In a homogeneous item set, these variants already exist, as all items are of the same type. Two mental rotation

tasks, for instance, will differ in the number of rotations they require, and nothing more. In heterogeneous item sets, in contrast, two items will differ in the features that characterize the task, making it difficult to pinpoint which item characteristic might drive differences in item difficulty. The general idea behind the two new experimental validation strategies is thus to deliberately manipulate individual characteristics of existing items. These manipulations are such that from the construct definition it can be expected that either the manipulated item is easier or harder than the original one (*change*-approach), or taps a different construct (*eliminate*-approach).

-- Insert Table 1 about here --

Four different analyses are required to investigate whether the two manipulations affected task performance as expected (Table 1) and will be described more detailed using the example of an ICT skills item (Figure 1).

-- Figure 1 about here --

To solve this item, the test-taker has to decide for each e-mail in his e-mail inbox whether it is relevant for a new colleague. If the user decides for relevance, he needs to forward the e-mail to the address that is provided in the instruction. The crucial aspect in this task is whether the third e-mail is identified correctly as a hoax e-mail that should not be forwarded.

2.1. Change-approach

The *change*-approach is based on the construct representation approach, in which item characteristics are related to item difficulty. But here, these characteristics are not identified for all items, but changed by developing a variant for a particular item where exactly this characteristic is changed. *Change* refers to a change of item-specific task characteristics that are assumed to evoke the construct that is supposed to cause differences in the test score. In

terms of the information-processing paradigm, *change* refers to a change of the cognitive process. The task solution should be easier or harder depending on the direction in which the processes are changed. A *change*-variant of the example item (Figure 1) can be created through changing the easiness to detect the third e-mail as a hoax e-mail. This aspect is crucial to the item as it requires ICT-specific evaluation skills. Since the presumed author of this e-mail is a rather trustworthy source, namely a colleague, the trustworthiness can be decreased in the *change*-variant by introducing an unknown author (a mailing list), potentially to the effect that the e-mail is read and evaluated more critically. If indeed the authorship serves as a criterion for evaluating e-mails, this item variant should be easier.

These considerations have two implications (cf. Table 1) for the functioning of changed items. First, depending on the nature of the change, the changed item should be easier or harder than the original. Second, the relations to other constructs should not be affected, as despite being easier or harder to perform, the cognitive processes required by an item (e.g., evaluating the e-mails) stay the same.

Previous studies already varied task characteristics in homogeneous item sets, for instance in matrices tasks. They followed predefined construction rules across all items and the purpose was for instance item writing (Hornke & Habon, 1986). In a matrices task, all items belong to the same item type because each item asks the test taker to identify a missing peace by applying different rules (e.g. addition). Although, the type of rules to be applied may differ across items, still each item is characterized by the requirement to apply one or more rules. We thus describe such item sets as homogeneous. We see the difference and innovation of the *change*-approach in that it can be also applied to heterogeneous item sets and that the purpose is in first line for validation but not for constructing new items. Other studies which were concerned with validation and also had to deal with heterogeneous item sets, related instead of task characteristics expert ratings, for instance regarding the cognitive

demands, to item difficulty (e.g. Watermann & Klieme, 2002). With the *change*-approach we suggest to manipulate those task characteristics (e.g. trustworthiness in the given example), which are assumed to require these cognitive demands. Since these manipulations can be made for every item separately, it does not matter how close and homogeneous items are to each other and whether comparable task characteristics can be found across items. Furthermore, while rating the cognitive demands delegates the validation process to the experts, manipulating task characteristics involves the test-taker stronger into the validation process. Similar as in earlier approaches of item manipulations (e.g., Hornke & Habon, 1986), also more than one *change*-manipulation could be possibly applied to one item, since items can be made easier or harder and also the degree of manipulation can vary.

2.2. Eliminate-approach

The *eliminate*-approach is based on investigating the nomothetic span. The relation to other variables being part of the assumed nomological network is evaluated for eliminate and original items. It is important to note that this approach is not primarily meant to investigate whether the relation exists as predicted by a nomological network, but goes further and compares the relations for manipulated and original items in order to investigate whether a change in the task characteristics affects the relation to other variables as expected. *Eliminate* refers to the elimination of all task characteristics that represent the construct, that is supposed to cause individual differences in the test score. Described in terms of the information-processing paradigm, elimination refers to the entire removal of the need to perform a specific cognitive process. *Eliminate*-items were created through elimination of the requirement to apply higher order ICT-skills involving judgement and decision. Thus, *eliminate*-items only required test-takers to perform basic operations, such as clicking buttons. Through this, presumably the nature of the targeted construct was changed. In the

example item, the correct e-mail that needed to be forwarded was already mentioned in the instruction. The last sentence of the instruction (Figure 1) "Now check your e-mails and forward important e-mails to Caro" was modified into "Now check your e-mails and forward the e-mail of Emma Martin to Caro". Again, these considerations have implications for the likely functioning of *eliminate*-items, as compared to the original item they were derived from. First, the probability of solving the item should be increased, if the requirement of performing a specific cognitive operation is removed from the item. Second, other than *change*-items, the correlations of eliminate-items to other variables should be affected, as removing the requirement to perform a specific cognitive process from an item will by definition change the nature of the construct assessed by the item.

We see the advantage of the eliminate-approach in constructing item-variants that lack the targeted skill dimension to investigate whether, besides item difficulty, the measured construct changes. This might seem to be not reasonable on the first sight, since these items can obviously not be used in further assessments. However, generally speaking correlation does not imply causation. Thus, for instance even if a computer-based reading test showed a strong correlation with a paper-based reading test, but not with a test of ICT skills, there would be always interpretations other than the intended (e.g., there is a common underlying ability, that "causes" the performance in both the computer and paper-pencil test of reading). In contrast, when *eliminate*-items are being administered to subjects randomly, the changes in item difficulty can be causally attributed to the manipulations in the items. Thus, the *eliminate*-approach allows to challenge seriously (Kane, 2013, p.15) the assumption that correlations of test scores with related variables are caused by the assumed skill dimension. It might seem to be trivial that a correlation changes once the targeted skill dimension is eliminated, but it is not trivial in items that are very complex and require for instance also some navigation or reading skills to read the instruction. For example, if the relation of test

scores from the ICT skills items to ICT related variables changes by eliminating the evaluation process from the item, it is supported that the relation was indeed caused by the required evaluation process.

2.3. Comparing the two approaches

How do the two strategies, *eliminate* and *change*, relate to each other? On a conceptual level, both manipulations differ in how they affect the cognitive processes while solving an item. The *change*-approach only affects the targeted cognitive process gradually (how difficult the evaluation process is), by making this easier or harder. The *eliminate*-approach, however, would eliminate all targeted cognitive processes (no evaluation process is required) that belong to the targeted skill dimension. This is why even if the evaluation of the hoax e-mail became rather easy, there will be still some evaluation skills needed in a *change*-item to treat this email correctly, but not in an *eliminate*-item. Thus, *eliminate*-manipulations can only lead to easier items, because cognitive processes are removed from the solution process, while *change*-manipulations can change difficulties in both directions and in different intensity. As a consequence, only one *eliminate*-manipulation can be carried out per item, while several manipulations are possible for *change*-items.

At second, the manipulations are carried out addressing different part of the items: *Change*-manipulations are carried out by changing task characteristics within the item, for instance the author of an e-mail, while *eliminate*-manipulations are carried out by adding information to the instruction. This is why an *eliminate*- and a *change*-item will never be the same although they may both decrease item difficulty.

And finally, they differ regarding the effect they are intended to have on construct-related variables. The *change*-manipulation leads to items that are intended to measure still the same construct and differ only in their difficulties, while the *eliminate*-manipulation leads

to items that should not measure anymore the same construct. As a consequence, *change*-items can be also used for eventual testing since they should measure the same construct, while *eliminate*-items cannot. Such *change*-variants can be useful for assessing specific samples, for instance regarding age or skill level, or for adaptive tests, where items with difficulties across the whole ability range are needed.

2.4. Theoretical and practical gains of the experimental approaches

One advantage of these procedures compared to correlational approaches (e.g. investigating the nomological network) is that potentially (several) confounding variables must not all be added to the validation design (although this of course comes at the price that the manipulated items need to be included in the assessment). With these procedures, only one construct-related variable can be used to investigate at first whether the expected relation actually exists (convergent evidence), and also at second and third, whether the relation to *change*- and *eliminate*-items changes or not.

When *change*- or *eliminate*-items are being administered to subjects randomly, the changes in test scores can be causally attributed to the changes in the items. By these means, the *change*- and *eliminate*-approach also add to the validity argument by addressing the cognitive processes that are assumed in a given item (AERA, APA, & NCME, 2014).

We consider especially the combination of both approaches as promising. The results of the *eliminate*- and *change*-approach strengthen each other: A relation that changes by manipulating task characteristics (*eliminate*-approach) supports that it is not trivial that the relation to *change*-items is the same after manipulating task characteristics, and vice versa. As positive side-effect, both approaches require considering the validation strategies already in the process of item development, which can be beneficial if the changes are already planned together with the item construction also for the original versions of the items.

Additionally, producing item variants takes only low effort once the original item is developed. This is not negligible, since implementing authentic items on a computer can be rather effortful and make feasible validation strategies even more important.

3. Applying the experimental approaches to the construct 'ICT skills'

3.1. Construct representation of ICT skills

In this research, we apply the *change-* and *eliminate-*approaches to validation to a test of ICT skills: ICT skills form a prototypical instance of a competence that is so broadly defined it can hardly be measured using a homogeneous set of items.

Different conceptualizations of ICT skills focus on different skill levels, such as basic computer skills (Goldhammer, Naumann, & Keßel, 2013), cognitive skills when using ICT (Eshet-Alkalai & Chajut, 2010), or the interplay of different levels of skills in one task (van Deursen & van Dijk, 2009). We focus on higher-order skills. Thus, we do not target basic ICT tasks that can be routinely performed on the basis of a pre-defined sequence of clicks. Rather, we target skills in such a way that they involve components of judgement and decision making (see the example item). A test measuring basic ICT skills might present test-takers with an e-mail, and then requiring them to enter a given address in the address field of some e-mail client, find, and click the "forward"-button. In contrast, higher-order ICT skills as addressed here would include a decision about whether a given e-mail should be forwarded to a given person or number of persons in a given situation. These decisions should be based on previous experiences, because experiences with ICT seem to determine skills (Eshet-Alkalai & Chajut, 2010). The decisions should be also based on knowledge specific to the ICT domain (henceforth "technical knowledge"), which is part of several ICT conceptualizations (Fraillon & Ainley, 2010; International ICT Literacy Panel, 2002; van Deursen & van Dijk, 2009). In our example, identifying the third e-mail correctly as a hoax

(cf. Figure 1) requires not only reading skills to understand the purpose of the e-mail but also evaluation skills in order to decide not to follow the call in the e-mail to forward. This e-mail is sent by a colleague who might be regarded as a trustworthy source (cognitive authority; Rieh, 2002). For this decision, higher-order ICT specific skills are needed that are based on knowledge and experience about typical markers of spam.

3.2. Developing a heterogeneous item set

In ICT environments, tasks can pose widely different cognitive challenges, or require different cognitive operations. For instance, a task might require a person to either access, manage, integrate, evaluate, or create information (International ICT Literacy Panel, 2002). In addition, the environment in which a task occurs can differ widely across tasks, and employ tools such as spreadsheets, browsers, e-mail clients, text-processors, etc. Thus, through the combination of ICT task and various environments, items are even within one cognitive operation heterogeneous. For instance, evaluate tasks may not only require to consider information regarding the author but other criteria of truth as well (Rieh, 2002). But also regarding the relevance of websites (Pfaff & Goldhammer, 2011), or the estimated value of information (Whittaker & Snider, 1996). If these different aspects of evaluating information, besides the other information tasks are included into the test, comparable and quantifiable criteria cannot even be found within all evaluate items. Although the construct is measured by heterogeneous items, we still assume that the construct of ICT skills is needed to solve all these items (i.e., assumption of one-dimensionality).

3.3. Hypotheses

Following the general steps for the *change*-approach, we expected the following (cf. Table 1): Changing task characteristics has an effect on item difficulty in the intended

direction (Hypothesis 1a). Moreover, the *change*-manipulation will not affect the effect of person covariates in changed items compared to original items (Hypothesis 1b).

Following the general steps for the *eliminate*-approach, we expected the following: *Eliminate*-items are easier than original items (Hypothesis 2a). Furthermore, by applying the *eliminate*-manipulation, the effect of person covariates in *eliminate*-items will be changed compared to original items (Hypothesis 2b).

3.4. Method

3.4.1. Sample

Both item manipulations were embedded in a calibration study of the ICT skills test. A sample of $N = 983$ (51% male, 46 % female, 3% not specified) was assessed. Participants were between 14 and 16 years ($M = 15.21$, $SD = 0.57$) and from 34 German schools from two federal states in Germany (Baden-Württemberg and Rheinland-Pfalz). Eleven schools belonged to the highest track (Gymnasium), and 23 schools to lower tracks.

3.4.2. Measures of person variables

The initial item pool consisted of 70 items which were implemented in a simulation environment by means of the CBA ItemBuilder (Rölke, 2012). The simulated applications in most items are browsers, e-mails, file managers, text processing software, spread sheet and presentation software (cf. Figure 1). Items were scored dichotomously. Behavior that could not classified as being definitely right or wrong was treated as neutral and did not count for the final score. For the given example we dealt with this in the following way: Three e-mails (first, third and fifth) should not be forwarded, the fourth e-mail has to be forwarded, and for the second e-mail both solutions are treated as correct. If the test-taker decided for one of the e-mails wrongly, the item was scored as incorrect (0), otherwise as correct (1). For the 70 original items, a one-dimensional Rasch model was fitted using TAM (Kiefer, Robitzsch, & Wu, 2016). Item-infits ranged between 0.87 and 1.11 and item-outfits between 0.67 and 2.18.

Two items were excluded from all analyses because of insufficient item-fit. The reliability of the model with 68 items was .70. The 42 original items that were manipulated and used for analyses had an average proportion of correct answers of $M = .47$ ($SD = .26$; $Min = .04$, $Max = .93$) and were thus from the whole range of difficulties.

As construct-related variables, technical knowledge and the frequency of ICT use were included. A subscale of the Computer Literacy Inventory (INCOBI-R; Richter, Naumann, & Horz, 2010) was used that assesses declarative computer knowledge with 20 multiple-choice items. Scores were computed by a total mean of correct answers ($M = .39$, $SD = .16$, $\alpha = .68$) and z-standardized for data analyses.

To assess ICT use, we asked students to estimate the frequency of seven specific activities in ICT environments in their daily lives. These activities were adapted from the PISA ICT Familiarity questionnaire (OECD, 2013) and assumed to represent such activities that have to be performed also in the test. These are how often they read and write e-mails, search for information for leisure or for school, read texts, create presentations and calculate for mathematics. We used a 4-point likert-scale with response categories "never", "several times a month", "several times a week", and "daily or almost daily". The variable "ICT use" represents the mean of those seven relevant activities ($\alpha = .73$) and was z-standardized.

3.4.3. Item manipulations

To create *change*-items, 40 items were selected from the 70 items. They were selected to be distributed across the five ICT-skills aspects access, manage, integrate, evaluate, and create nearly equally in order to have a good representation of the ICT specific aspects (access, manage and create: eight items; integrate seven items; evaluate: nine items). Whether items were made easier or harder was very specific to the items. If an item could be assumed to be hard on theoretical grounds, the item was changed to become easier. Correspondingly, if an item could be assumed to be easy on theoretical grounds, it was changed to become

harder. From the 40 items, 30 items were intended to become easier and 10 items intended to become harder. Since younger persons struggle with evaluation tasks (Eshet-Alkali & Amichai-Hamburger, 2004), the example item was assumed to be already comparatively hard (the hoax e-mail was sent from a trustworthy person). Thus, we opted for a change that presumably would decrease the item's difficulty by changing the author in a less trustworthy mailing list. Likewise, a possibility of increasing the item's difficulty could have been to introduce an even more trustworthy person as the sender of the spam-e-mail (e.g. a supervisor).We applied only one manipulation per item, since this allowed us to use the available testing time to vary rather more items instead of varying one item in two different directions, which is important in the face of a heterogeneous item pool. For a smaller or less heterogeneous item pool, an even more ambitious procedure could include giving some test takers an easier item-variant (e.g. author is a mailing list) and other test-takers the harder item-variant (e.g. author is a supervisor). To create *eliminate*-items, 20 items were selected and stripped of any requirements to apply higher-order ICT skills involving judgement and decision making. These 20 items were equally distributed across the five ICT aspects.

Excluding two items led finally to 38 *change*-items (29 easier, 9 harder) and 18 *eliminate*-items for analyses.

3.4.4. Procedures

The assessment consisted of two parts (cf. Table 2), while each part took about one hour. Before the students started with the test, all received a tutorial to become familiar with the simulated environment. Then, students were assigned randomly to the different booklets, and worked in the first part either on original items ($n = 773$) or *change*-items ($n = 210$), but never on both. In the second part of the assessment, *eliminate*-items and questions for technical knowledge and ICT use were administered. From those students who worked in part one on original items, 220 students received *eliminate*-items in the second part of the test.

Due to a balanced design of original items in the first part, regarding information tasks, applications, and estimated time intensities (Wenzel et al., 2016). Some of those students ($n = 173$) received in part one of the test already an original version of an *eliminate*-item. Although we minimized this number of overlapping items, this happened on average for three items per person ($M = 2.98$, $SD = 2.09$). As a consequence, the answer on the corresponding *eliminate*-item was not used for analyses, in order to avoid that a second presentation of the same item could have affected the results. Questions regarding ICT use were administered to all students, while a few students ($n = 284$) did not receive the technical knowledge questions.

-- Insert Table 2 about here --

This design was chosen to ensure at first a well-balanced design for the 70 original items for calibration, by administering as many original items as possible to one student and by balancing items regarding content and time-intensity. We decided to administer change-items parallel to original items in the first part to avoid for motivational reasons that students worked in both parts on demanding and time-intensive ICT skills items. Besides, administering change-items alike eliminate-items in the second part would also have led again to a second presentation of item-variants. This could not be avoided due to strong overlaps in original, change- and eliminate-variants.

3.4.5. Data analyses

Generalized linear mixed models (GLMM; De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx, & Partchev, 2011; Wilson, De Boeck, & Carstensen, 2008) available in the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014) were used for all hypotheses. With GLMM we refer to a more general analysis framework allowing for IRT models being explanatory on item side (cf. LLTM; Fischer, 1973) but also doubly explanatory including both item and person covariates (latent regression LLTM), as

well as including an error component on item side (LLTM+e; Janssen, Tuerlinckx, Meulders, & DeBoeck, 2003) as random effect (cf. Wilson et al., 2008). In a GLMM, the probability to solve an item correctly is expressed by the logit of the probability P to solve the item correctly, which can be explained by fixed effects, denoted by the Greek letter "β", and random effects, denoted by the Latin letter "b". Equation 1 contains the model that was applied for all analyses. The effect β_0 represents an overall intercept. If also group-specific intercepts β_{0k} are modeled to compare the original to the manipulated items, β_0 refers only to the reference group of original items. To relate each manipulated item to the corresponding original item, the corresponding items were treated as equal but differed in their group membership g, which led also to a group specific random item intercept, b_{0ig}, representing the (residual) item easiness. The random person intercept b_{0p}, represents (residual) person ability. Since students were nested in schools, we also included a random intercept for schools, b_{0s}. A fixed effect β_{0k} was modeled to investigate whether the manipulated items became indeed easier and harder (k) compared to the original items (β_0). For Hypotheses 1b and 2b, additional fixed effects were modeled to investigate whether the manipulated items differ in their relation to the person covariate (v), β_{vk}, from the relation of the original items to the person covariate, β_v.

$$ ln\left[\frac{P_{pi}}{1-P_{pi}}\right] = \beta_0 + \sum_{k=1}^{K} \beta_{0k}X_{(p,i)0k} + \beta_v X_{(p,i)v} + \sum_{k=1}^{K} \beta_{vk}X_{(p,i)v}X_{(p,i)0k} + b_{0ig} + b_{0p} + b_{0s} \quad (1) $$

GLMMs include the negative difficulty as item parameter, that is, higher and positive values describe a higher probability of successful task solution and thus easier items. The easiness of an item is represented by the fixed intercept for all items and item-specific deviation from this.

3.5. Results

3.5.1. Change

In line with Hypothesis 1a (Table 3), *change*-manipulations worked in both directions. Items that were intended to become easier were indeed easier than the original items ($\beta = 0.54$, $p < .001$) and items that were intended to become harder were indeed harder than the original items ($\beta = -0.90$, $p < .001$).

-- Insert Table 3 about here --

To investigate the influence of construct-related variables (Hypothesis 1b), a model was estimated for the 38 original items and their counterparts, the manipulated *change*-items. The results of Hypothesis 1b (Table 4) indicated that both ICT-related variables are as expected positively related to the probability of success in the original items (technical knowledge: $\beta = 0.29$, $p < .001$; ICT use: $\beta = 0.09$, $p = .006$). Also in line with the hypothesis, the easier *change*-items did not differ from this relationship (technical knowledge: $\beta = -0.09$, $p = .196$; ICT use: $\beta = -0.04$, $p = .611$), and the harder items differed only for technical knowledge into the positive direction (technical knowledge: $\beta = 0.21$, $p = .037$; ICT use: $\beta = 0.10$, $p = .337$), which means that the probability of success in these items were even stronger related to technical knowledge than the original items.

-- Insert Table 4 about here --

Results from the *change*-approach support, that the *change*-items were as intended easier or harder, and seemed to measure still the same construct.

3.5.2. Eliminate

Supporting Hypothesis 2a (Table 5), the *eliminate*-items were indeed easier than their original counterparts ($\beta = 1.45$, $p < .001$).

-- Insert Table 5 about here --

To investigate whether manipulations affected the measured construct (Hypothesis 2b; Table 6) the relation of construct-related variables to the probability of success was estimated for original items, again for the 18 original items and their *eliminate*-counterparts.

-- Insert Table 6 about here --

The results of Hypothesis 2b indicated, that both ICT-related variables were as expected positively related to the probability of success in the original items (technical knowledge: β = 0.25, p < .001; ICT use: β = 0.11, p = .017). In line with the hypothesis, the relation to ICT use differed for the *eliminate*-items indeed from this relation (β = -0.24, p < .001), however, the relation to technical knowledge did not differ for the *eliminate*-items (β = -0.00, p =.961).

Results from the *eliminate*-approach support, that *eliminate*-items were as intended easier and seemed to measure a (partly) different construct, since the relation to ICT use changed but not the relation to technical knowledge.

4. Discussion

In the present paper, we introduced two novel approaches to validate test items, *eliminate* and *change*. These approaches allow to relate task characteristics to test scores and can be applied even to heterogeneous items sets, as they are more the rule than the exception in "modern educational assessments" (Baumert, et al., 2009, p.166), and also used in the assessment of 21st century skills. Such constructs are often assessed in a contextualized way, which makes the items rather complex. The suggested approaches are particularly useful to investigate whether test scores represent indeed differences in the targeted processes. Using ICT skills as an example, results indicated that *changing* item-specific task characteristics in the items affected item difficulty in the intended direction. These changes did not affect the to-be-measured construct, since the relations to technical knowledge and ICT use were not affected by the manipulation. Only the probability of success in those items that were

manipulated to be harder were even stronger explained by technical knowledge compared to the original items. *Eliminating* the targeted skill dimension led to easier items and affected the to-be-measured construct partly, since the relation to ICT use is different for *eliminate*-items but not the relation to technical knowledge. In the following, we discuss these results and interpretations especially regarding technical knowledge and what can be gained from these approaches for test development.

4.1. Consequences for test score interpretation of the ICT skills test

How can we interpret the results regarding the targeted test score interpretation? Although the relation to the construct-related variables did as expected not change by applying the *change*-manipulation, items that were manipulated to become harder had an even stronger relationship to technical knowledge (cf. Table 4). This might be because task characteristics requiring already technical knowledge (e.g. knowledge about spam e-mails) are, beside other task characteristics, likely starting points for manipulations. That technical knowledge is even more decisive in items that were manipulated in harder direction does not necessarily speak against the targeted construct interpretation. In the example item for instance (Figure 1), knowledge about spam is required to identify typical markers of spam and to decide correctly not to forward the hoax e-mail. If for instance a hoax e-mail was sent by a more trustworthy author (e.g. a supervisor instead of a colleague), knowledge about hoax e-mails is likely to be even more decisive for a correct task solution. Test score interpretation would have been rather called into question if these harder items were less related to technical knowledge than the original items. Besides, the relation to ICT use supports that test scores from both *change*-groups can be interpreted in a similar way as the original items, thus, that changing the difficulty of those items did not change the construct.

Against our expectation, the relation to technical knowledge was not affected by the *eliminate*-manipulation. Since we assume that technical knowledge is an integral part of higher-order ICT skills, eliminating higher-order ICT skills should also have affected the relation to technical knowledge. Thus, we expected that when items do not require applying such knowledge, for instance about hoax emails as it is the case in *eliminate*-items, this relation should be affected. What does this mean for the test score interpretation? At first, that we manipulated not those task characteristics in *eliminate*-items that cause the relation to test scores from technical knowledge. Thus, we did either manipulate the wrong task characteristics, or technical knowledge scores represent not, or not only, knowledge that we assumed to be relevant for higher-order skills. That the relation to technical knowledge could be even increased by the *change*-manipulation, supports that the identified task characteristics were somehow related to technical knowledge. This is why we should have a closer look to what test scores from technical knowledge might represent and what we understood by technical knowledge.

From the understanding in our study, technical knowledge plays a double role in the construct we focus on, in ICT skills. At first as integral part of higher-order ICT skills, but also on a lower level as part of basic ICT skills as they are required for navigating (Goldhammer et al. 2013). Finding for instance a forward button (cf. Figure 1) may require some knowledge about e-mail environments. The scale we used for technical knowledge might possibly not differentiate between technical knowledge that is related to lower and higher-order skills. Thus, even if it is possible to increase the relation to technical knowledge by manipulating knowledge as in the *change*-manipulation, it might be not possible to eliminate this relationship completely because navigation might still require to some extent technical knowledge as it is represented by the scale.

Possibly, we underestimated the role of technical knowledge for merely interacting with ICT environments. However, this does not minor the relevance of the validation approach, but rather implies that technical knowledge as chosen variable was not appropriate. However, the relation to ICT use supports that test scores from *eliminate*-items cannot be interpreted in the same way as the original items, which is strongly supported by the even negative relation to ICT use for the *eliminate*-items. Thus, we changed the construct at least partly with the *eliminate*-manipulation.

Further item manipulations could help to investigate the role of technical knowledge. One manipulation could contain, for instance, to keep only higher-order processes in the item by eliminating the navigation from the item. This could be reached by presenting for instance screenshots of the items and to compare then the relation of technical knowledge to the probability of success in such items to the relation of technical knowledge to the probability of success for original items. If the relationship changes, technical knowledge is indeed required for navigation. Taken together, we learned that the entanglement of different levels of skills involved in CBA items is not trivial and that specific attention should be paid to validity of test scores assessed with complex items as they are used for instance in educational assessments.

4.2. Deeper analyses and implications for test developers

The suggested approaches can provide valuable and additional information regarding the single items and task characteristics for test developers. Although the used method allows investigating at first only the average change of item difficulties due to the applied manipulation, deeper analyses can be conducted.

Firstly, it can be analyzed whether the changes were differently effective in different items by referring to the variance in items above the average effect. This can be reached by

comparing a model with item-specific adaptation to the *change*-effect (with random effect) to a model without item-specific adaptation (no random effect). Such analyses are especially useful if there was no average change in item difficulty, to investigate for instance whether only a few items did not change, or whether all changes were too small. Item-specific adaptations to the average intercepts can indicate which items changed most or less, or even in the wrong direction. If an item did not change, this can be for instance because the manipulated task characteristic was not at all used by the test-takers as assumed (e.g. evaluation processes were not performed at all), because the change was not effective and did not affect the evaluation process, or because the original item was already very easy or hard for the test-takers.

Secondly, it can be helpful to group items regarding task characteristics, for instance, items that require similar evaluation processes, if the number of items per task characteristic is sufficient and the selected task characteristics for the grouping are meaningful. This allows analyzing whether indeed all groups of task characteristics affected item difficulty. These deeper analyses can help reconsidering theoretical assumptions and indicators (cf. Kane, 2013, p.40).

4.3. Conclusion

Using experimental strategies for test score validation, if successful, can support the plausibility of test score interpretation, because the targeted test score interpretation is challenged. Although the process of validation depends on the test and the construct, the *eliminate-* and *change*-approaches provide a general strategy for validation that can be transferred to other constructs and contexts. This is especially the case in the area of educational measurement, where broad constructs are used. These constructs are often assessed by means of computers, allowing the simulation of authentic settings. This may lead

the same time to heterogeneous item sets, where current validation approaches cannot be applied to. The two suggested strategies combine experimental techniques with the recent concept of validation. They provide a concrete and systematic approach for implementing the modern understanding of validity. For this reason they can be regarded as a valuable tool assuring a theory-based operationalization of constructs through test items.

Acknowledgements

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). Standards for Educational and Psychological Testing. Washington: AERA, APA, NCME.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, URL http://CRAN.R-project.org/package=lme4.

Baumert, J., Lüdtke, O., Trautwein, U. & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*, 165-176.

Binkley, M., Erstad, O., Herman, J., Raizen, R., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17 – 66). Dordrecht: Springer.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Caissie, A. F., Vigneau, F., & Bors, D. A. (2009). What does the Mental Rotation Test measure? An analysis of item difficulty and item characteristics. *Open Psychology Journal, 2*, 94-102.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Eshet-Alkali, Y. & Amichai-Hamburger, Y. (2004). Experiments in digital literacy. *CyberPsychology & Behavior, 7*, 421-429.

Eshet-Alkalai, Y., & Chajut, E. (2010). You Can Teach Old Dogs New Tricks: The Factors That Affect Changes over Time in Digital Literacy. *Journal of Information Technology Education, 9*, 173–181.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374. doi:10.1016/0001-6918(73)90003-6

Fraillon, J., & Ainley, J. (2010). The IEA international study of computer and information literacy (ICILS). Retrieved from http://www.researchgate.net/profile/John_Ainley/publication/268297993_The_IEA_Int ernational_Study_of_Computer_and_Information_Literacy_%28ICILS%29/links/54eba 4330cf2082851be49a9.pdf

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). Preparing for life in a digital age. Springer-Verlag GmbH.

Goldhammer, F., Naumann, J. & Keßel, Y. (2013). Assessing Individual Differences in Basic Computer Skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*, 263-275.

Greiff, S., Krkovic, K., & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56*, 83-103.

Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital

text. *Computers in Human Behavior, 55*, 486–500.

http://doi.org/10.1016/j.chb.2015.09.042

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der

Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten [Using the

prediction of item difficulties for construct validation and model-based proficiency

scaling]. *Psychologische Rundschau, 63,* 43-49. doi:10.1026/0033-3042/a000109

Hornke, L. F. & Habon, M. W. (1986). Rule-based item bank construction and evaluation

within the linear logistic framework. *Applied Psychological Measurement, 10,* 369-380.

International ICT Literacy Panel (2002). Digital Transformation: A Framework for ICT

Literacy. Princeton, NJ. Retrieved from

http://www.ets.org/research/policy_research_reports/publications/report/2002/cjik

Janssen, R., Tuerlinckx, F., Meulders, M., & DeBoeck, P. (2003). A hierarchical IRT model

for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*,

25, 285-306. doi:10.2307/1165207

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of

Educational Measurement, 50*, 1-73.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test Analysis Modules*. R package version

1.16-0. http://CRAN.R-project.org/package=TAM

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment:

Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

OECD (2011), PISA 2009 Results: Students on Line: Digital Technologies and Performance

(Volume VI). http://dx.doi.org/10.1787/9789264112995-en

OECD (2012), Literacy, Numeracy and Problem Solving in Technology-Rich Environments:

Framework for the OECD Survey of Adult Skills, OECD Publishing.

http://dx.doi.org/10.1787/9789264128859-en

OECD (2013), PISA 2012 Assessment and Analytical Framework: Mathematics, Reading,

    Science, Problem Solving and Financial Literacy, OECD Publishing.

    http://dx.doi.org/10.1787/9789264190511-en

OECD (2014), PISA 2012 Results: What Students Know and Can Do – Student Performance

    in Mathematics, Reading and Science (Volume I, Revised edition, February 2014),

    PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

Pfaff, Y., & Goldhammer, F. (2011, September). *Measuring individual differences in ICT*

    literacy: Evaluating Online Information. Talk presented at the 14th Biennial

    Conference of the European Association for Research on Learning and Instruction

    (EARLI), Exeter, United Kingdom.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation

    for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Richter, T., Naumann, J. & Horz, H. (2010). Das Inventar zur Computerbildung (revidierte

    Fassung). *Zeitschrift für Pädagogische Psychologie, 24*, 23-37.Rieh, S. Y. (2002).

    Judgment of information quality and cognitive authority in the Web. *Journal of the*

    American Society for Information Science and Technology, 53, 145-161. doi:

    10.1002/asi.10017

Rölke, H. (2012). The item builder: A graphical authoring system for complex item

    development. In T. Bastiaens & G. Marks (Eds), *Proceedings of World Conference on*

    E-Learning in Corporate, Government, Healthcare, and Higher Education (pp. 344 –

    353). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/41614.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In

    pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna

    (Eds.), *Handbook of Test Development* (pp. 329-347). Mahwah, New Jersey: Lawrence

    Erlbaum Associates, Inc.

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior, 65*, 100-106.

van Deursen, A. J., & van Dijk, J. A. (2009). Using the Internet: Skill related problems in users' online behavior. *Interacting with Computers, 21*, 393–402. doi:10.1016/j.intcom.2009.06.005

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal Of Psychological Assessment, 18*, 190-203. doi:10.1027//1015-5759.18.3.190

Wenzel, S. F. C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F., Naumann, J., & Horz, H. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) (CavE-ICT) [Computer-based, adaptive and behaviorrelated assessment of information and communication-related competencies (ICT skills)]. In, Forschung in Ankopplung an Large-scale AssessmentsBMBF (Hrsg.) (pp. 161e180). Bonn, Berlin: BMBF.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91-120). Göttingen: Hogrefe.

Whittaker, S., & Sidner, C. (1996, April). Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 276-283). ACM.

Figure 1. Example of an ICT skills test item.

Table 1

Analyses for the experimental approaches.

	Indicators for task performance	
Manipulation	Item difficulty	Relation to construct related variables
Changing (specific aspects)	H1a: Easier or Harder (than original items)	H1b: Same pattern (as original items)
Eliminating (a whole skill dimension)	H2a: Easier (than original items)	H2b: Different pattern (than original items)

Table 2

Design of the study.

Group	Part 1		Part 2	$N = 983$
1		Original Items	Eliminate-Items + ICT Use +Technical Knowledge	220
2	Tutorial	Original Items	B R E A K ICT Use + Technical Knowledge	269
3		Original Items	ICT Use	284
4		Change-Items	ICT Use + Technical Knowledge	210

Table 3

Hypothesis 1a: Probability to solve change-items (38) compared to original items (38).

Parameters		β	*SE*	*z*	*p*
Fixed					
Intercept (Original)		-0.13	.27	-0.48	.629
Change Items: Intended easier (29)		0.54	.15	3.56	<.001
Change Items: Intended harder (9)		-0.90	.26	-3.50	<.001
Random					
Variance (person)	0.38				
Variance (school)	0.16				
Variance (item)	Intercept (Original)	2.48			
	Change	0.55			

Note. Model: value ~ Intended Change + (group | item) + (1 | person) + (1 | school); persons = 973, schools = 34, number of observations (persons x answered items) = 16260.

Table 4

Hypothesis 1b: Estimated effects of the technical knowledge and ICT use. Change-items are compared in their relation to the two variables (Change (easier/harder): Variable…) to the relation of the original items to the two variables (Variable (Original)…).

Parameters			…Technical Knowledge				…ICT Use			
			β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Fixed										
Intercept (Original)			-0.12	.27	-0.44	.662	-0.12	.27	-0.45	.650
Change Items: Intended easier (29)			0.54	.15	3.56	<.001	0.56	.15	3.65	<.001
Change Items: Intended harder (9)			-0.98	.25	-3.91	<.001	-0.89	.26	-3.47	<.001
Variable (Original)…			0.29	.04	6.75	<.001	0.09	.03	2.72	.006
Change (easier): Variable…			-0.09	.07	-1.29	.196	-0.04	.07	-0.51	.611
Change (harder): Variable…			0.21	.10	2.08	.037	0.10	.10	0.96	.337
Random										
Variance (person)		0.37					0.38			
Variance (school)		0.11					0.15			
Variance (item)	Intercept (Original)	2.60					2.48			
	Change	0.50					0.54			

Note. Models: value ~ Intended Change * variable + (group | item) + (1 | person) + (1 | school); Model for Technical Knowledge: persons = 681, schools = 34, number of observations (persons x answered items) = 12166; Model for ICT Use: persons = 948, schools = 34, number of observations (persons x answered items) = 15952.

Table 5

Hypothesis 2a: Probability to solve eliminate-items (18) compared to original items (18).

Parameters			β	*SE*	*z*	*p*
Fixed						
Intercept (Original)			-0.24	.32	-0.76	.446
Eliminate Items			1.45	.31	4.72	<.001
Random						
Variance (person)		0.52				
Variance (school)		0.25				
Variance (item)	Intercept (Original)	1.65				
	Eliminate	1.58				

Note. Model: value ~ group + (group | item) + (1 | person) + (1 | school); persons = 762, schools = 34, number of observations (persons x answered items) = 7944.

Table 6

Hypothesis 2b: Estimated effects of the technical knowledge and ICT use. Eliminate-items are compared in their relation to these variables

(Eliminate: Variable…) to the relation of the original items to these variables (Variable (Original)…).

Parameters		…Technical Knowledge				…ICT Use			
		β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Fixed									
Intercept (Original)		-0.29	.31	-0.93	.355	-0.25	.32	-0.78	.437
Eliminate Items		1.51	.32	4.70	<.001	1.45	.31	4.72	<.001
Variable (Original)…		0.25	.06	4.05	<.001	0.11	.05	2.38	.017
Eliminate: Variable…		-0.00	.08	-0.05	.961	-0.24	.07	-3.34	.001
Random									
Variance (person)		0.58				0.51			
Variance (school)		0.19				0.24			
Variance (item)	Intercept (Original)	1.59				1.65			
	Eliminate	1.73				1.58			

Note. Models: value ~ group * variable + (group | item) + (1 | person) + (1 | school); Model for Technical Knowledge: persons = 479, schools = 34, number of observations (persons x answered items) = 5632; Model for ICT Use: persons = 742, schools = 34, number of observations (persons x answered items) = 7782.