Engelhardt, Lena; Naumann, Johannes; Goldhammer, Frank; Frey, Andreas; Wenzel, S. Franziska C.; Hartig, Katja; Horz, Holger

Convergent evidence for validity of a performance-based ICT skills test

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
European journal of psychological assessment 36 (2020) 2, S. 269-279

Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Wenzel, F.C., Hartig, K., & Horz, H. (2020). Convergent evidence for the validity of a performance-based ICT skills test. *European Journal of Psychological Assessment, 36(2),* 269-279.
https://doi.org/10.1027/1015-5759/a000507

Convergent evidence for the validity of a performance-based ICT skills test

Article Type: Original Article

Word Count: 6995

Tables

Figures

Abstract

The goal of this study was to investigate sources of evidence of convergent validity supporting the construct interpretation of scores on a simulation-based ICT skills test. The construct definition understands ICT skills as reliant on ICT-specific knowledge as well as comprehension and problem-solving skills. On the basis of this, a validity argument comprising three claims was formulated and tested. (1) In line with the classical nomothetic span approach, all three predictor variables explained task success positively across all ICT skills items. As ICT tasks can vary in the extent to which they require construct-related knowledge and skills and in the way related items are designed and implemented, the effects of construct-related predictor variables were expected to vary across items. (2) A task-based analysis approach revealed that the item-level effects of the three predictor variables were in line with the targeted construct interpretation for most items. (3) Finally, item characteristics could significantly explain the random effect of problem-solving skills, but not comprehension skills. Taken together, the obtained results generally support the validity of the construct interpretation.

Skills related to information and communication technologies (ICTs) are variously described as 21st century skills (Binkley et al., 2012), survival skills (Eshet-Alkalai, 2004), and key competencies for lifelong learning (European Parliament and the Council, 2006). Their importance has turned them into an object of assessments (Ferrari, Punie, & Redecker, 2012). Variables that tend to be considered as a source of evidence for the convergent validity of ICT skills scores are demographic in nature, and include gender, socio-economic status, self-reports on the use of ICT, and self-efficacy (Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016, p.33). We suggest that one reason for this is because theoretical assumptions about the associated skills are rather vague; no clearly defined construct definition including the skills required and their interplay exists. Many conceptualizations consider ICT skills to be a mixture of technical proficiency and other skills that are not exclusive to ICT contexts. These additional skills are described and labelled in various ways, and include reasoning, metacognitive skills, critical thinking, reading, problem-solving, and numerical skills (Calvani, Cartelli, Fini, & Ranieri, 2009; International ICT Literacy Panel, 2002). Fraillon and Ainley (2010) label these skills more abstractly, as “conventional literacies”. Given that they are not exclusively related to ICT contexts, it is not surprising that such constructs are not conventionally considered sources of evidence for convergent validity. However, as this study focuses on the validity of the construct interpretation, our first goal is to specify what these additional skills are, their interplay, and how they relate to ICT skills in order to formulate a testable validity argument.

A second challenge in considering convergent sources of validity evidence is the wide variety of ICT tasks that could potentially be addressed in test items. Nearly every information task can be performed using ICTs. Potential ICT tasks might require evaluating the trustworthiness of information or managing a numerical table. As a consequence, the required skills might vary with different tasks. For example, different tasks might require

reading skills to a greater or lesser extent. Such variation alone does not necessarily pose a threat to the validity of test score interpretation. However, if a few items require no ICT-specific skills, only conventional skills, one could question whether scores on these items reflect ICT skills at all. Thus, the second goal of this study is to apply a task-based approach to investigate whether construct interpretation is also valid on the item level.

This study focuses on a simulation-based ICT skills test of 15-year-old students' ability to handle everyday ICT tasks. The goal of this study is to provide sources of evidence of convergent validity supporting the construct interpretation of the test scores. Our argument-based approach applies an understanding of validity in accordance with that of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, & NCME, 2014), in which validity refers not to the test itself but rather the test score interpretation.

ICT Skills

We begin by describing the targeted construct interpretation before discussing relevant skills and their interplay. These then serve as a basis for formulating testable hypotheses.

Construct Interpretation of Test Scores

Performing tasks in an ICT environment requires various types of skills. Whereas every ICT task requires, at a minimum, interacting with an ICT environment via mouse clicks, touch gestures or typing, some tasks also require making decisions on the basis of complex considerations. An example of an ICT task requiring basic skills (cf. Goldhammer, Naumann, & Kessel, 2013) in operating technology would be forwarding an email. A task requiring higher-order ICT skills would also contain, in addition to forwarding the email, making a decision on whether the email should be forwarded or not. Making such a decision requires

detecting and taking into account spam markers or email credibility criteria. Thus, the ability to perform this type of decision-making should be captured in the associated test score. ICT tasks incorporating these types of decisions include choosing relevant books in a library database or deciding between two language courses based on the content of their websites. The higher-order ICT skills needed to make these decisions should be reflected in scores on performance-based ICT skills tests.

Cognitive Processes Assumed to be Involved in Solving ICT Tasks

We argue that ICT-specific skills, such as ICT-specific knowledge, are necessary, but not sufficient, to solve ICT tasks; additional skills are involved when making decisions that necessitate higher-order ICT skills. These additional skills are required to process and understand information and interact with the environment. We assume that the cognitive processes involved in understanding the information presented in the environment consist of comprehending textual (e.g. words) and graphical (e.g. images) information (Schnotz, 2005). We assume that the cognitive processes involved in navigating the environment, in performing and organizing the steps required to reach a defined goal are similar to the processes involved in problem-solving (Simon & Newell, 1971). These two skills, comprehension and problem-solving, are typically involved in working with ICTs and have both been investigated in the context of digital environments (Naumann & Sälzer, 2017; Organisation for Economic Co-operation and Development [OECD], 2012). The following sections provide a theoretical description of reading comprehension and problem-solving processes in order to identify task characteristics that might increase the need for reading comprehension or problem-solving skills in ICT tasks.

Reading

According to Kintsch's (1998) construction-integration (CI) model, text comprehension starts with lower-level processes for processing letters and words and then requires building a propositional representation of the text's contents (textbase model). Further processes then integrate prior knowledge in order to construct a situation model. Comprehension processes are primarily affected by the quality with which individual words are represented by the reader (Perfetti, 2007). In line with this, the quality of lexical representations on the phonological, orthographic, and meaning levels can predict text comprehension on an inter-individual level (Richter, Isberner, Naumann, & Kutzner, 2013). On an intra-individual level, more frequent words can be assumed to be better represented (e.g. Just & Carpenter, 1987; Kaakinen & Hyönä, 2010; White, Warrington, McGowan, & Paterson, 2015). In the addition to the frequencies of individual words, the syntactic complexity of a text determines how difficult it is to comprehend. One widely-used indicator of syntactic complexity was established by Flesch (1948) and captures the average length of both words and sentences in a text (Coke & Rothkopf, 1970; England, Thomas, & Paterson, 1953). In line with the notion that sentences comprising more syntactic phrases will be longer and more difficult to process (e.g. Graesser, Hoffmann, & Clark, 1980; Marton, Schawartz, & Brown, 2005; Schindler, Richter, Isberner, Naumann, & Neeb, 2018), texts with longer sentences ought to be more difficult to comprehend. Finally, the amount of information represented is also assumed to matter in reading tasks (OECD, 2012, p.24), described for instance by number of sentences or words (Mesmer & Hiebert, 2015; Walkington, Clinton, Ritter, & Nathan, 2015).

Problem-solving

We assume that the cognitive processes needed to navigate an ICT environment, to perform and organize the different steps required to reach a defined goal, involve problem-solving processes (Simon & Newell, 1971). Simon and Newell describe problem-solving as taking place in a problem space in which problem solvers have to find their way by selecting

different operators to reach a certain goal state. The size of the problem space, and thus the number of operators that have to be selected, might be related to the difficulty of a given problem. Problem difficulty can be described by referring to the number of variables (e.g. in microworlds; Stadler, Niepel, & Greiff, 2016), elements and transformations (e.g. geometric analogies; Embretson, 1983), or steps (e.g. Tower of Hanoi problem; Spitz, Webster, & Borys, 1982). The amount and diversity of behavior required to solve a problem in a technology-rich environment describes that problem's intrinsic complexity (OECD, 2012, p.50). The number of navigational steps has also been used empirically as a component of item difficulty (Naumann, Goldhammer, Rölke, & Stelter, 2014).

ICT-specific knowledge

We assume that ICT-specific knowledge can help guide comprehension and problem-solving processes, such as seeking out specific information in a database or figuring out how to forward an email in a new environment. We focus on higher-order ICT skills, which are required to make correct decisions in ICT tasks and are based on conventional comprehension and problem-solving skills as well as on ICT-specific knowledge.

Whether and to what extent these conventional skills are needed depends strongly on the characteristics of the task. For example, the presence of words and information units determines whether comprehension processes are evoked, whereas the length and structure of a task from start to end state determines whether problem-solving processes are elicited. A task using spreadsheet or presentation software might contain less text and thus require lower-order comprehension skills than a task in a browser environment. Searching for a document on the computer in a deep, complex folder structure may lead to longer solution paths and thus require problem-solving skills to a greater extent. ICT-specific knowledge should be required in every ICT task, although the difficulty of the task and thus the required

knowledge might vary. Knowledge about spam markers might be less widespread and harder to apply than knowledge about the functioning of email inboxes.

The relevance of ICT-specific and conventional skills for solving a given ICT task is assumed to depend on task difficulty. ICT-specific skills are assumed to become increasingly important with item difficulty. Conventional skills, in contrast, are assumed to be an important prerequisite for dealing with typical ICT tasks. Tasks that are primarily hard because they require higher levels of comprehension or problem-solving skills are not the focus of this study, because these reflect differences in ICT skills to a minor extent. Rather, we focus on ICT tasks that are primarily hard because more advanced ICT knowledge has to be applied and integrated into the task solution.

Convergent Sources of Evidence for the Construct Interpretation

In this section, we define convergent sources of evidence. We start off by following the classical nomothetic span approach (Embretson, 1983), which assumes relations on the test score level. In the standards for educational and psychological testing (AERA, APA, & NCME, 2014), such an approach is known as validity based on relations to other variables. Support for the interpretation of a test score with respect to a certain criterion requires investigating test-criterion relations; for instance, the association between the test score and the frequency of using technical devices. However, the present study deals with construct interpretation. Therefore, we provide sources of evidence of convergent validity by investigating relations to constructs that are assumed to require similar cognitive processes. More specifically, we seek to provide for the construct interpretation that higher test scores represent differences in ICT-specific knowledge as well as differences in conventional skills (as we focus on higher-order ICT skills). The intended construct interpretation would not be supported if one of those three constructs was not related to ICT skills test scores.

The construct interpretation of the test score is supported if unique positive effects on solving ICT skills items can be found for all three predictors, that is, ICT specific knowledge, and problem-solving, and comprehension skills (Hypothesis 1).

Even if the probability of success in ICT skills items can be positively predicted on average by construct-related knowledge and skills, this might not necessarily be true for every single item due to item-specific composites of task characteristics. Task characteristics determine the type and extent of knowledge and skills that are required for solving the task. Variation in the effect of a given skill across items might depend on how the items were designed (e.g., the amount and complexity of text presented to the test-taker). Variation in effects does not call the item design into question as long as it is in line with the intended construct interpretation. Thus, ICT-specific knowledge should influence the probability of success on any ICT skills item. Additionally, problem-solving and/or comprehension skills should play a role. Extending the classical nomothetic span approach, the relations between construct-related knowledge and skills variables should also be investigated at the item level in order to clarify whether the pattern of effects fits the intended construct interpretation. Although we expect that both conventional skills will be related to scores on the ICT skills test on the test level, individual items might be related to either problem-solving (e.g., due to navigation requirements) or reading comprehension (due to reading requirements) only. As some items might require one of the conventional skills to only a limited extent, the item score interpretation is valid according to Figure 1, if each item requires ICT-specific knowledge and at least one of the conventional skills (thus ensuring that higher-order skills are present). Accordingly, the construct interpretation would not be supported for an item if either (1) ICT-specific knowledge had no effect or (2) reading comprehension *and* problem-solving had no effect (see this interplay in Figure 1 in the Electronic Supplementary Material

[ESM 1]). However, if at least one of the conventional skills had an effect, the construct interpretation would be supported.

The construct interpretation of the item score is supported if positive effects on solving ICT skills items can be found for ICT-specific knowledge and problem-solving, or ICT specific knowledge and comprehension skills, or ICT specific knowledge, and problem-solving, and comprehension skills (Hypothesis 2).

Varying effects of construct-related knowledge and skills could be caused by varying item characteristics. The construct representation approach (Embretson, 1983) quantifies item characteristics that should evoke the processes underlying the construct interpretation. A relation between these indicators and item difficulties then supports the targeted construct interpretation. Item characteristics evoking reading comprehension and problem-solving processes can be identified for ICT skills items. An item's reading load can be assumed to indicate the required reading comprehension processes, whereas the number of steps in the solution process indicates the required problem-solving processes. It is more difficult to identify quantifiable task characteristics for ICT-specific knowledge (AUTHOR, 2017). The task characteristics related to comprehension and problem-solving skills are not alone sufficient to support construct interpretation, as they only capture conventional demands, not ICT-specific demands. However, they are nevertheless important, as they can support the test score interpretation if they actually moderate the effects of conventional skills. As a result, they can lend further support to the notion that reading and problem-solving processes occur in the items as expected.

The construct interpretation of test scores is supported if the strength of the positive effects of comprehension and problem-solving skills depends on quantifiable item characteristics that determine how strongly these cognitive processes are evoked (Hypothesis 3).

Method

Sample and Procedures

The sample consisted of $N = 269$ 15-year-old German students ($M = 15.29$, $SD = 0.68$, $Min = 14$, $Max = 17$) roughly equally split between males (52%) and females (46%; rest not specified). Schools in two federal states, Baden-Württemberg and Rhineland-Palatinate, were asked if they would be willing to participate. Thirty-four volunteering schools equipped with suitable computer equipment were then selected to participate in the study. Eleven of the selected schools were of the highest German track (Gymnasium). Most participating students were in Grade 9 (74%), with the rest in tenth grade. Prior to testing, the students' parents provided written declarations of consent that their children were allowed to participate in the study.

The assessment consisted of two parts, each of which lasted about one hour. Before beginning the test, test-takers received a tutorial to familiarize them with the simulated computer environment. As it was not possible to assign every item in the first part of the assessment to every student due to time constraints, students were randomly assigned to four different versions of the test, leading to missing data by design (missing completely at random). The different test versions were balanced according to the content of the ICT tasks, the ICT application used, and the estimated time required to complete the individual items, and thus assumed to be interchangeable. This led to different numbers of items in the different test versions (i.e. 30-33 items). Students completed $M = 26.18$ ($SD = 6.03$; 1st quartile = 23; 3rd quartile = 32) items on average. Students could refuse to work on an item and navigate to the next item on their own (omissions were treated as incorrect). The test-level time restriction of one hour also led to not-reached items (treated as not administered). Out of the total number of items that could potentially be completed, 2.4% were omitted and

17.4% were not reached. In the second part of the assessment, test-takers received questions assessing their ICT-specific knowledge, the reading comprehension test, and the problem-solving test. Data from the 256 students who completed all tests was used in the analyses.

Person Variables

The ICT skills items were developed in accordance with the International ICT Literacy Panel's (2002) conceptualization, which distinguishes among ICT tasks involving accessing, managing, integrating, evaluating, and creating information. They were implemented in a simulated computer environment by means of the CBA ItemBuilder[1] (Rölke, 2012), a tool for creating dynamic and interactive tasks for computer-based assessments (CBA). Applications such as browsers, e-mail clients, and file managers were simulated. The performance-based items were scored dichotomously. After excluding six items due to item fit and differential item functioning (AUTHOR, 2016), 64 items were selected for the test.

Problem-solving was assessed with seven items from the Complex Problem-solving Scale (Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013). For each item, scores for knowledge acquisition (expected a posteriori (EAP) reliability: .77) and knowledge application (EAP reliability: .75) were extracted and fitted in a two-dimensional two-parameter logistic item response model using the R package TAM (Kiefer, Robitzsch, & Wu, 2014). Only the knowledge acquisition score was used for analyses as it is conceptually closer to the assumed role of problem-solving in an ICT context.

As to our knowledge no test covering the comprehension of both textual and graphical elements exists, we decided to apply a well-established time-limited reading comprehension test. Test-takers needed to complete gaps in a text by choosing the most appropriate of three

[1]More recent information about the CBA ItemBuilder can be found on this web page: https://tba.dipf.de/en/infrastructure/software-development/cba-itembuilder/cba-itembuilder-1?set_language=en (last accessed 8.3.2018)

presented words within a time limit of four minutes. According to the authors of this test, the comprehension score has a retest-reliability of $r = .87$ (German Reading Speed and Comprehension Test; Schneider, Schlagmüller, & Ennemoser, 2007). Comprehension scores were constructed from the sum of correct solutions (given a score of 2), including a penalty for incorrectly solved items (given a score of -1), in accordance with the manual, and were standardized for analyses.

ICT-specific knowledge was assessed using a subscale of the Computer Literacy Inventory (Richter, Naumann, & Horz, 2010) assessing theoretical computer knowledge. This scale consists of 20 multiple-choice questions (each had four answer alternatives) about different terms concerning computers. The sum of correctly answered questions was counted and standardized for analyses. For this sample, Cronbach's alpha was $\alpha = .68$.

Task Characteristics

Three different indicators were computed to capture the reading demands generated by the presented textual information for each of the 64 items (including the instructions): First, we estimated the frequency of each individual word in the text and averaged these frequencies for each item separately. We used the SUBTLEX database based on German subtitles (Brysbaert et al., 2011), because measures extracted from subtitles turn out to be particularly predictive of lexical decision performance (Soares et al., 2015). The average frequencies across all items were $M = 3766.52$ ($SD = 766.96$, $Min = 2199.33$, $Max = 5876.77$). Second, the Flesch index was calculated for each item using the R package koRpus (Michalke, 2017) to describe each text's readability on a scale of 0 to 100 ($M = 57.58$, $SD = 9.97$, $Min = 33.80$, $Max = 83.29$), with higher values indicating easier texts. Third, we use the number of words as indicator for the amount of information presented ($M = 235.90$, $SD = 250.33$, $Min = 45$, $Max = 1815$).

The number of required user interactions was counted for each item on the basis of the expected correct solution to describe item difficulty resulting from problem-solving demands (*M* = 6.00, *SD* = 3.43, *Min* = 1, *Max* = 16). As iterative behavior is not assumed to increase the problem-solving skills required, only unique behaviors were counted. For example, opening five emails in a row was only counted as one required interaction. Thus, this indicator combined the number of steps with the diversity of behavior (cf. OECD, 2012, p.50). All variables were standardized for analyses.

Data Analyses

We applied generalized linear mixed models (GLMM; Wilson, De Boeck, & Carstensen, 2008) using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014). This method allows relations between covariates and item responses to be investigated as fixed and random effects. Fixed effects assume a constant relation across all items, whereas random effects allow for variation, assuming different relations for different items. Random effects of the GLMM are assumed to be normally distributed (Wilson et al., 2008). The probability of solving an item correctly was expressed by the logit of the probability of one person (p) solving one item (i) correctly (P_{pi}) and higher values can be interpreted in terms of item easiness. The lme4 package excludes missing observations and observations with missing values for any variable from the model (cf. Bates et al., 2014). All data analyses were based on *N* = 6207 observations.

$$ln\left[\frac{P_{pi}}{1-P_{pi}}\right] = \beta_0 + \sum_{v=1}^{V} \beta_{1v} X_{(p,i)v} + \mathrm{b}_{0\mathrm{i}} + \mathrm{b}_{0p} + \mathrm{b}_{0s} \quad (1)$$

Equation 1 describes the model used to analyze Hypothesis 1, which was then extended for the other analyses. An overall intercept (β_0) as well as random effects across items (b_{0i}), persons (b_{0p}), and schools (b_{0s}) were modeled and the person variables v (problem-solving,

reading comprehension, computer knowledge) were included as fixed effects (β_{1v}). To analyze Hypothesis 2, a full random effects model (Model 2) was applied to allow for variation in the relations of the person variables across items. That is, the effects of the three variables were allowed to vary across items and to correlate with item easiness, as well as among each other. Whether these variations contributed significantly to model fit was investigated by comparing this model to the model for Hypothesis 1. Hypothesis 2 refers to the variation in the variables' explanatory value for each item. Because the tested model does not provide an inference statistic for each variable and each item, we developed the following rule to determine whether a variable has explanatory value for a particular item. First, we transformed the item-specific effects on the logit scale of Model 2 into probabilities (cf. Equation 2) and compared the probability of solving a specific item for a theoretical person with a high value (+ 1.96 SD) to that of a theoretical person with a low value (-1.96 SD) in a particular person variable (e.g. computer knowledge). In Equation 2, γ_i denotes the item-specific effect of the person variable based on the fixed and random effects in Model 2 ($\beta_{1v} + b_{1vi}$), and δ_i the item easiness based on the fixed and random effects ($\beta_0 + b_{0i}$) in Model 2. Note that in Equation 2, the ICT skills of the person and school as well as all other person covariates are assumed to be 0, corresponding to their respective means. Then, we used the following rule of thumb: If the probability of a correct solution is less than 5% higher for a person with a high value in a person variable (e.g. computer knowledge) compared to a person with a low value on that variable, the variable is not considered to have explanatory value for that item.

$$P(X_{pi} = 1)_{high/low\ skilled} = \frac{exp^{(\gamma_i*(\pm 1.96)+\delta_i)}}{1 + exp^{(\gamma_i*(\pm 1.96)+\delta_i)}} \quad (2)$$

To test Hypothesis 3, interaction terms between the variables and task characteristics were added to investigate whether the relation between the person variables and ICT skills

was stronger for items with stronger reading and problem-solving demands. One indicator was used for problem-solving (Model 3) and three indicators for reading comprehension (Models 4a-4c). The original input and output for all four models can be found in the Electronic Supplementary Material (ESM 2).

Results

According to *Hypothesis 1*, all three person variables had unique positive effects on task success across all ICT items (Table 1, Model 1; computer knowledge: $\beta = 0.18$, $p < .001$, reading comprehension: $\beta = 0.17$, $p < .001$, problem-solving: $\beta = 0.31$, $p < .001$). Including these variables explained 35.7 % of person variance and 71.0% of school variance.

--- Insert Table 1 about here ---

The full random effects model (Model 2) fitted the data better than the model without variation (χ^2 (9) = 20.09, $p = .017$), indicating varying effects across items (see Table 1 in the Electronic Supplementary Material [ESM 3]). The variation in the effects of all three variables across items is visualized in Figure 1, sorted by item difficulty. The random effects are displayed as variation from the fixed effect. Problem-solving was positively related to task success for all items. The relation between computer knowledge and task success was around zero for a few easy items, whereas a zero relation with reading comprehension was found for a few medium and very difficult items. Whereas the effect of computer knowledge tended to be higher for more difficult items ($r = -.63$), the effect of conventional skills tended to be higher for easier items (RC: $r = .50$; PS: $r = .16$). To address Hypothesis 2, the explanatory values of the person variables by item were determined on the basis of the item-specific effects from Model 2 and the differences for high versus low skilled persons in the three person variables were calculated for each item (see Table 2 in the Electronic Supplementary Material [ESM 4]).

--- Insert Figure 1 about here ---

According to *Hypothesis 2*, all items for which computer knowledge was not explanatory (given the 5% rule) should be removed with the option of being revised. This was the case for five easy items and for the most difficult item. In addition, all items for which both reading comprehension and problem-solving skills were not explanatory should be removed. Although reading comprehension had no explanatory value for a couple of items, problem-solving skills were explanatory for all but for the hardest three items, meaning that only the construct interpretation of these three items was called into question additionally to the five easy items. As these three items were only solved correctly by a few persons, it is not surprising that none of the examined skills could increase the probability of task success by at least 5%. Thus, considering these results together with the high and positive random effects of computer knowledge and problem-solving skills on those harder items (cf. Figure 1), we argue for keeping those items. However, the five easy items for which reading and problem-solving skills were indeed explanatory, but not computer knowledge, should be removed with the option of being revised.

Item characteristics assumed to evoke reading and problem-solving processes were included in the analyses (*Hypothesis 3*; Tables 2 and 3). As expected, the intrinsic complexity of a task interacted positively with the effect of problem-solving (Model 3; $\beta = 0.11$, $p = .017$). Contrary to our expectations, no indicator for reading load interacted with the effect of reading comprehension (Models 4a-4c; word frequency: $\beta = 0.02$, $p = .683$; readability: $\beta = -0.02$, $p = .642$; number of words: $\beta = 0.02$, $p = .679$).

--- Insert Table 2 and 3 about here ---

Discussion

Across all items, all three constructs - reading comprehension skills, problem-solving skills, and computer knowledge - predicted task success positively (Hypothesis 1). Problem-solving was the strongest predictor of task success across all items (fixed effects; Figure 1) and was also more predictive on the item level than the ICT-specific skill, computer knowledge, for most items (random effects; Figure 1). We suggest that this might be due to the operationalization of the tests. The problem-solving test required dynamic interaction with a surface, which might have caused the higher correlations between problem-solving and the performance-based ICT skills items. The computer knowledge test, in contrast, required only limited interaction with the environment and can even be administered on paper. Taken together, the three variables explained a substantial amount of person variance (35.7%) in solving ICT skills items.

On the whole, the construct interpretation was supported (Hypothesis 2), as computer knowledge and at least one of the conventional skills were required for most items. However, the item-specific effects of computer knowledge were around zero for a few easy items and the explanatory value of computer knowledge was less than 5% for those items. Thus, the construct interpretation might be not valid for these items, because computer knowledge did not serve to differentiate between persons. We have therefore excluded these items with the option of being revised. Although the explanatory value for reading comprehension was also low in a couple of items, because problem-solving was still predictive for those items, the construct interpretation remains valid.

We assumed that the relations between conventional skills and ICT skills were due to similar cognitive processes triggered by task characteristics. As expected, the number of user interactions required for item solution explained the relation with problem-solving, whereas the relation with reading comprehension could not be explained by any of the indicators (Hypothesis 3). Two explanations are possible. First, one could question the validity of the

test score interpretation, because it is possible that the correlation between reading comprehension and ICT skills was not due to similar processes performed in ICT skills and reading tasks. If the association with reading was caused by third variables rather than task characteristics, the construct definition would not be true for the ICT skills items, as they would not involve reading processes. However, the variation in the association with reading comprehension across items speaks against this interpretation, as it indicates that the association with reading depends on item properties. Thus, a second – more plausible – explanation calls into question the reading indicators as a reliable indicator for reading load, as they were not at all related to item difficulty, with the exception of word frequency (cf. Table 3). We assume that the indicators could not represent reading demands, because not all text and words on a page necessarily need to be read to solve the items correctly. Future research could therefore analyze which words or elements are processed in test items (for instance, via eye tracking methods) in order to identify necessary elements for task solution and to establish a better indicator. In contrast to the reading indicators, the problem-solving indicator did indeed reflect steps that had to be performed to solve the item correctly.

Considering all these results together, the validity of the construct interpretation of the test score was generally supported for everyday ICT tasks performed by 15-year-old students in Germany. Only a few easy items should be excluded and re-checked, as ICT-specific skills were not decisive for solving these items in this sample. It should further be taken into account that the lme4 package used for testing GLMMs does not provide statistics on absolute model fits. Instead, we addressed the issue of model fit solely by comparing the relative fits of models with and without random effects (Bolker et al., 2009) to assess whether the full random effects model was needed to represent the data.

Conclusion

In this study, we extended the classical nomothetic span approach and investigated relations to other variables on the item level as well as interaction effects with task characteristics. Significant variation in the relations across items supported the appropriateness of this approach. Variables selected for the collection of validity evidence are often criterion variables such as self-reports or the usage of ICTs. However, this study drew upon skills that are needed during the task solution process, namely comprehension and problem-solving skills in addition to computer knowledge. Considering those skill variables allows for the validation of construct interpretations in contrast to criterion-referenced test score interpretations (cf. AERA, APA, & NCME, 2014). Moreover, we created a bridge between a domain that tends to be rather operationally defined, namely ICT skills, and well-established and well-studied conventional constructs. Thus, we could overcome the shortcomings of recent conceptualizations by providing detailed suggestions on the nature and reasons for the associations with underlying skills and knowledge. This allowed us to collect validity evidence for the construct interpretation of test scores from the ICT skills test based on relations to other constructs.

Electronic Supplementary Material

ESM 1. Figure 1 (1ESM_Figure1.tif).

This figure (adapted from AUTHOR, 2016 (p. 164)) shows the definition of ICT skills and describes the different skills that are assumed to be relevant for solving ICT tasks. The dark grey area describes how ICT skills are understood in this study.

ESM 2. Output (2ESM_Output.pdf).

This file shows the input and output of the models 1-4.

ESM 3. Table 1 (3ESM_Table1.tiff).

This table shows the results of the full random effects model (model 2) with person variable effects varying across items, co-variating with item easiness and among each other in order to approach Hypothesis 2.

*ESM 4.*Table 2 (4ESM_Table2.tiff).

This table shows the difference in the probability of task solution in ICT-skills items for a person with high (+1.96 SD) compared to a person with low (-1.96 SD) computer knowledge (CK), reading comprehension (RC), and problem-solving skills (PS) skills. Differences lower than 5% are grey shaded.

References

AUTHOR, 2016

AUTHOR, 2017

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bates, D., Maechler, M., Bolker, B., & 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.

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.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127-135. doi:10.1016/j.tree.2008.10.008

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*, 412-424. http://dx.doi.org/10.1027/1618-3169/a000123

Calvani, A., Cartelli, A., Fini, A., & Ranieri, M. (2009). Models and instruments for assessing digital competence at school. *Journal of e-Learning and Knowledge Society-*

English Version, *4*, 183–193. Retrieved from http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/288/270.

Coke, E. U., & Rothkopf, E. Z. (1970). Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology, 54*, 208-210. doi:10.1037/h0029067

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197. http://dx.doi.org/10.1037/0033-2909.93.1.179

England, G. W., Thomas, M., & Paterson, D. G. (1953). Reliability of the original and the simplified Flesch reading ease formulas. *Journal of Applied Psychology, 37,* 111-113. doi:10.1037/h0055346

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia*, *13*, 93-107. Retrieved from http://www.openu.ac.il/Personal_sites/download/Digital-literacy2004-JEMH.pdf

European Parliament and the Council (2006). Recommendation of the European Parliament and the Council of 18 December 2006 on key competences for lifelong learning. *Official Journal of the European Union, L394*. Retrieved from http://www.alfa-trall.eu/wp-content/uploads/2012/01/EU2007-keyCompetencesL3-brochure.pdf

Ferrari, A., Punie, Y., & Redecker, C. (2012). Understanding digital competence in the 21st century: an analysis of current frameworks. In *21st Century Learning for 21st Century Skills* (pp. 79-92). Springer Berlin Heidelberg. doi:10.1007/978-3-642-33263-0_7

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233. doi:10.1037/h0057532

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_Intern

ational_Study_of_Computer_and_Information_Literacy_%28ICILS%29/links/54eba4330cf2082851be49a9.pdf.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment*, *29*, 263-275.

Graesser, A. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior, 19(2)*, 135-151.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of Complex Problem-solving: concept, implementation, and application. *Educational Technology Research and Development, 61*, 407-421.

International ICT Literacy Panel (2002). Digital Transformation: A Framework for ICT Literacy. *Educational Testing Service*. Princeton, NJ. Retrieved from http://www.ets.org/Media/Research/pdf/ICTREPORT.pdf.

Just, M. A., & Carpenter, P. A. (1987). Speed reading. In M. A. Just & P. A. Carpenter (Eds.), *The psychology of reading and language processing* (pp. 425–452). Newton, MA: Allyn & Bacon.

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1561–1566. http://dx.doi.org/10.1037/a0020693

Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: An R Package for Item Response Modelling.*

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Marton, K., Schwartz, R. G., & Braun, A. (2005). The effect of age and language structure on working memory performance. In B. G. B. L. Barsalou, & M. Bucciarelli (Eds.),

Proceedings of the XXVII. Annual Meeting of the Cognitive Science Society (pp. 1413-1418). Mahwah, NJ: Erlbaum.

Mesmer, H. A., & Hiebert, E. H. (2015). Third Graders' Reading Proficiency Reading Texts Varying in Complexity and Length:Responses of Students in an Urban, High-Needs School. *Journal of Literacy Research, 47*, 473-504. doi:10.1177/1086296x16631923

Michalke, M. (2017). koRpus: An R Package for Text Analysis. Available from https://reaktanz.de/?c=hacking&s=koRpus

Naumann, J., Goldhammer, F., Rölke, H. & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen [Successful Problem-solving in Technology Rich Environments: Interactions Between Number of Actions and Task Demands]. *Zeitschrift für Pädagogische Psychologie, 28*, 193-203. doi 10.1024/1010-0652/a000134

Naumann, J., & Sälzer, C. (2017). Digital reading proficiency in german 15-year olds: evidence from PISA 2012. *Zeitschrift für Erziehungswissenschaft, 20*, 585-603. doi:10.1007/s11618-017-0758-y

OECD (2012). *Literacy, Numeracy, and Problem-solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills,* OECD Publishing.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383.

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*)* [A Revised Version of the Computer Literacy Inventory]. *Zeitschrift für Pädagogische Psychologie*, *24*, 23-37.

Richter, T., Isberner, M.-B., Naumann, J. & Kutzner, Y. (2013). Lexical quality and reading comprehension in primary school children. *Scientific Studies of Reading*, *17*, 415-434. https://doi.org/10.1080/10888438.2013.764879

Rölke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012* (pp. 344-353). Chesapeake, VA: AACE.

Schindler, J., Richter, T., Isberner, M.-B., Naumann, J & Neeb, Y. (2018). Construct validity of a process-oriented test assessing syntactic skills in German primary school children. *Language Assessment Quarterly: An International Journal*. Advance online publication. doi: 10.1080/15434303.2018.1446142

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12: Lesegeschwindigkeits- und –verständnistest für die Klassen 6-12* [Reading Speed and Comprehension Test for Grades 6 to 12]. Göttingen: Hogrefe.

Schnotz, W. (2005). An Integrated Model of Text and Picture Comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–69). New York, NY, US: Cambridge University Press.

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

Simon, H. A., & Newell, A. (1971). Human problem-solving: The state of the theory in 1970. *American Psychologist, 26*, 145–159. http://dx.doi.org/10.1037/h0030806

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., . . . Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *The Quarterly Journal of Experimental Psychology, 68*, 680-696. doi:10.1080/17470218.2014.964271

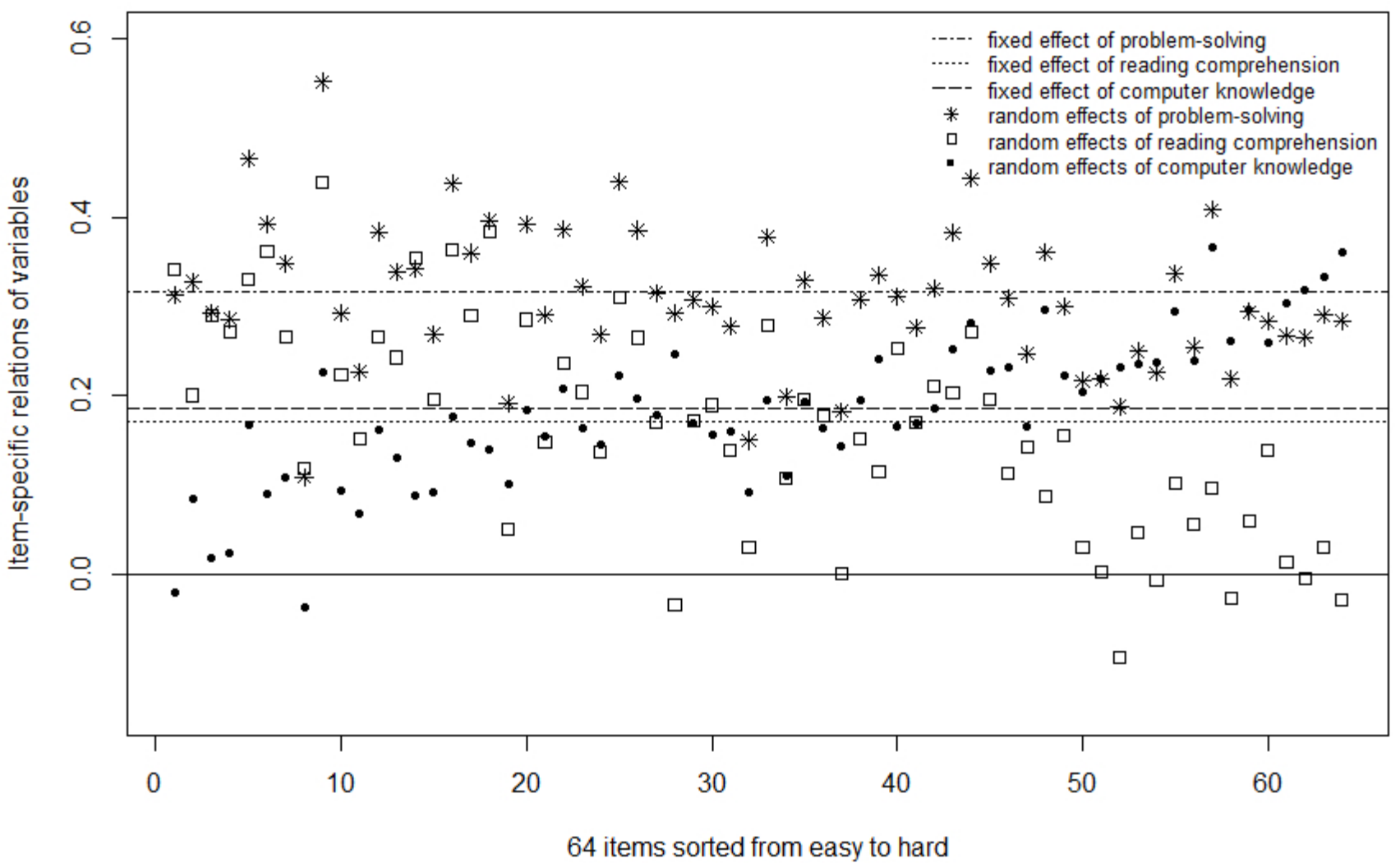
Spitz, H. H., Webster, N. A., & Borys, S. V. (1982). Further studies of the Tower of Hanoi problem-solving performance of retarded young adults and nonretarded children. *Developmental Psychology, 18*, 922-930. doi:10.1037/0012-1649.18.6.922

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior, 65*, 100-106. https://doi.org/10.1016/j.chb.2016.08.025

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology, 107*, 1051-1074. doi:10.1037/edu0000036

White, S. J., Warrington, K. L., McGowan, V. A., & Paterson, K. B. (2015). Eye movements during reading and topic scanning: Effects of word frequency. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 233-248. doi:10.1037/xhp0000020

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.


fixed effect of problem-solving
fixed effect of reading comprehension
fixed effect of computer knowledge
random effects of problem-solving
random effects of reading comprehension
random effects of computer knowledge
Item-specific relations of variables
0.0
0.2
0.4
0.6
0
10
20
30
40
50
60
64 items sorted from easy to hard