Original Article

Assessing Skills of Everyday Technical Problem Solving

Marcus Schrickel[1,2], Jennifer Stemmann[3], Frank Goldhammer[2,4], and Carolin Hahnel[1,2]

[1]Faculty of Psychology, Ruhr University Bochum, Germany
[2]Centre for Technology-based Assessment (TBA), DIPF | Leibniz Institute for Research and Information in Education, Germany
[3]Technology and its didactics, University of Education Freiburg, Germany
[4]Centre for International Student Assessment (ZIB), Frankfurt, Germany

Abstract: With the ongoing digitalization of home automation and appliances, users often face situations in which they lack the information required to reach an intended goal when using an unfamiliar device. To solve such technical problems, users need to acquire and apply operational knowledge in a self-regulated way. Our study assesses skills in technical problem solving (TPS) and provides the first empirical validation of a score interpretation of a developed TPS test. The examined hypotheses concern the internal structure of the TPS test and the association of the item solution rate with device characteristics (complexity and intransparency) and person-by-device characteristics (the extent of system exploration, use of operating manuals, and device-specific prior knowledge). An ad hoc adult sample (n = 213) processed nine computer-simulated TPS units, each including a knowledge acquisition and knowledge application phase per device. As expected, the internal structure of the TPS test was one-dimensional. Intransparency, but not complexity, affected the item solution rate. Completeness of system exploration, use of operating manuals, and device-specific prior knowledge independently showed positive effects on item solution rates. Along with an evaluation of the developed TPS test, we discuss the role of the salience of device features for knowledge acquisition.

Keywords: computer-based assessment, problem solving, test development, home automation and appliances, lifelong learning

OPEN DATA OPEN MATERIALS OPEN CODE PREREGISTERED

Home automation and appliances, such as heating thermostats, washing machines, or telephones, are part of everyday life. With ongoing digitalization, those devices are evolving from analog towards digital interfaces, often requiring new or updated knowledge about how users can interact with provided structures to achieve a particular result (operational knowledge; Wirth, 2004). Simple or intuitive operating routines, such as the turning of an analog heating thermostat dial, are replaced by interactions that take place at a digitized user interface. Such changes in the device design are often accompanied by expanding their range of functions (e.g., setting up a precise desired temperature). Although users are expected to operate everyday technologies competently and independently (Lindqvist et al., 2018), they are rarely instructed or trained (see Tamas et al., 2021) and usually need to gather new or additional knowledge and establish operating routines on their own.

Situations in which a person operates a device but lacks the information required to reach an intended goal can be considered a technical problem (see Dörner, 1976). Technical Problem Solving (TPS) refers to skills of actively acquiring or activating operational knowledge and using it purposefully to solve such technical problems. Despite its relevance for everyday life, there is a lack of instruments assessing TPS. To this end, we developed such a test and examined the validity of its intended test score interpretation. We analyzed the test's psychometric properties and associations with device and person-by-device characteristics.

Problem Solving When Using Home Automation and Appliances

TPS describes individual skills of purposefully operating an unfamiliar device with the goal of transforming it from an initial state into a desired target state (Stemmann, 2016). To overcome barriers within this transformation process, a person needs to acquire and apply operational knowledge about the system. For example, a user might want to change the system time of a newly bought device, transferring its system from a default state into a state with the desired settings. Challenges might include the yet unfamiliar structure of the device menu, requiring the user to gain knowledge on how to navigate and operate the menu to adjust specific settings. Such interactions take place in a problem

https://doi.org/10.1027/1015-5759/a000847

space that contains the current state of a problem situation and all possible states into which the current situation can be transformed, with one or more states representing the target state (Wirth, 2004).

The degree to which a problem requires interaction is a distinguishing characteristic between static problem solving, in which all information is immediately available, and dynamic problem solving, in which information needs to be generated (Greiff et al., 2012). These interactions also differentiate *problems* from *tasks* (Dörner, 1976), referring to the role of non-routine and routine operations. When solving problems, individuals cannot rely solely on routine operations, as they initially lack the means to overcome the barriers to reach an intended goal, for example, due to an unfamiliar menu structure.

Technical problems can vary in their *complexity* and *intransparency* (see Funke, 2001). Regarding technical systems, Stemmann (2016) refers to complexity as a system's number of states, variables, and control elements describing the system. A device with a low number of menu items and manipulatable variables is less complex than a device with an extensive menu structure and many modifiers. While other definitions of complexity also include the relationship between variables (Stadler et al., 2019), Stemmann attributes those to the distinct feature of interconnectedness. Intransparency is the extent of system states, system variables, and their interconnections being temporally or permanently unknown (Stemmann, 2016). For example, a device with numerically labeled menu items is more intransparent than a device with semantically labeled menu items.

Acquiring and Applying Operational Knowledge in TPS

Before solving a technical problem, individuals need to acquire knowledge about the problem space. According to Wirth (2004), *knowledge acquisition* (exploring the problem space) and *knowledge application* (finding a solution) are two process components with different goals that are intertwined during the problem solving process. For complex problem solving (CPS), for example, Lotz, Scherer, Greiff and Sparfeldt (2017) point out that both process components occur intermittently during problem solving in natural settings, but are separated for assessment purposes. Accordingly, findings on the empirical distinctness of both process components are not always clear (e.g., Kretzschmar et al., 2017, find evidence for a one-factor structure for complex problem solving; Wüstenberg et al., 2012, find evidence for two empirically distinct but correlated dimensions). For TPS, empirical evidence supports the differentiation of both process components, as the performance in knowledge application phases correlated positively but moderately with indicators of individuals' knowledge acquisition (Stemmann, 2016).

There are different ways of acquiring knowledge about technical problems. Individuals may systematically explore a device to generate information about its states and variables (Stemmann, 2016) or read instructions from operating manuals, which introduces aspects of information problem-solving (Brand-Gruwel et al., 2009) into TPS (e.g., the need to identify appropriate terms to search for specific information). Generating specific operational knowledge will be guided by prior knowledge about devices and general device structures acquired in everyday life (see Klahr & Dunbar, 1988). The role of prior knowledge distinguishes TPS from other problem solving constructs, such as CPS, where domain-specific prior knowledge is considered an undesired influence (Wüstenberg et al., 2012) or even irrelevant (Süß & Kretzschmar, 2018), and advanced domain-specific problem solving, which requires professional training (e.g., Abele et al., 2012, studied problem solving of professional mechatronics and electronics technicians). Therefore, technical problems are located "in the middle" on a continuum between problems that can be solved independently of prior knowledge and highly domain-specific problems.

TPS is also closely related to other cognitive constructs. For example, although it can be assumed to be related to intelligence, TPS cannot be clearly attributed to a specific intelligence factor, such as fluid reasoning (the use of mental operations to solve novel problems) or comprehension knowledge (declarative and procedural knowledge; McGrew, 2009). TPS also covers aspects of information and communication technology (ICT) skills, particularly with regard to the operation of devices with digital interfaces (e.g., see Binkley et al., 2012). However, TPS has a broader focus on technical problems related to everyday technologies, including home automation and appliances.

Independent from knowledge acquisition, accuracy in the assessment of one's own problem solving skills may impact TPS performance. Accurate prospective metacognitive judgments (see Mihalca & Mengelkamp, 2020) can contribute to resource allocation when solving a problem (Dentakos et al., 2019). For example, an overconfident user might allocate fewer cognitive resources than required and fail to acquire the necessary operational knowledge to solve a technical problem due to poor exploration and manual use. In turn, underconfidence can result in resources being used for too long and to an excessive extent (Dentakos et al., 2019).

Current Study

When everyday devices present unfamiliar barriers due to technological innovation, users cannot rely on their previous operating routines and need to activate skills in dealing with technical problems (TPS). We developed a test to quantify and research TPS skills and behaviors in an adult population. Our primary goal was to focus on knowledge application in TPS and derive a test score that reflects the extent of a person's skill to purposefully apply operational knowledge when solving a technical problem.

A TPS test should provide testees with non-routine tasks, in which they can demonstrate their TPS skills by interacting with everyday devices. To this end, we adapted and refined device simulations by Stemmann (2016). Her test consisted of 15 computer-simulated everyday devices, covering five device types that share similarities in appearance, system behavior, and purpose (i.e., household appliances, home automation, multimedia devices, self-service machines, information and communication technology). System configurations (i.e., the extent of the device characteristics complexity, intransparency, dynamics, and connectivity) differed across devices but were not varied systematically. Each device provided testees with a knowledge acquisition phase, in which they could explore the device for 10 minutes, and a subsequent one-item knowledge application phase, in which they had to reach a particular goal state of the system within 5 minutes (ω = .71). However, Stemmann's test did not provide any operating manuals typical for everyday knowledge acquisition and the characteristics of the items and devices were confounded, as only one item is presented per device. Our adaption addressed these limitations.

Following the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), the present study is a first attempt to validate the construct interpretation of the score from the adapted TPS test that represents knowledge application as an individual TPS skill. For this purpose, we investigated (1) the psychometric properties of the adapted TPS test as well as (2) the effects of device characteristics and (3) person-by-device characteristics on item success. First, we psychometrically examined if all items require the same skill. As the TPS items should capture individuals' skills to apply operational knowledge to solve technical problems, we expected the internal structure of the TPS test to be one-dimensional (hypothesis H1). Second, we expected the device characteristics complexity (H2a) and intransparency (H2b) to be negatively associated with item success. To clearly attribute possible effects to device characteristics, we additionally controlled for the item characteristic of the number of actions required to solve an item correctly. This way, we ensure, for example, that a possible effect of complexity does not arise from many actions being required to solve an item correctly. Third, we expected the completeness of exploration (H3a) and manual use (H3b) as indicators of knowledge acquisition to be positively associated with item success. As prior knowledge potentially supports exploring the problem space (Klahr & Dunbar, 1988), we also expected it to positively impact TPS item success (H3c). Finally, as the degree of the prospective metacognitive judgment of one's problem solving skill may determine the allocation of cognitive resources (see Dentakos et al., 2019), we expected accurate judgments to positively impact TPS item success too (H3d). Additionally, explorative analyses concerned the effect of psychological immersion (i.e., the experience of being involved, absorbed, engaged, or engrossed; Lombard et al., 2009) as a control variable, associations between device characteristics and knowledge acquisition (post hoc), a possible interaction effect of exploration and manual use and participants' regular behavior in TPS. While examining H1 provides validity evidence concerning the internal test structure, examining the other hypotheses provides validity evidence based on response processes (H2a-b, H3a-b) and concerning relations to other variables (H3c-d) (see AERA, APA, & NCME, 2014).

Materials and Methods

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models.

If we use inferential tests, we report exact p values, effect sizes, and 95 % confidence or credible intervals.

Sample

We planned for a sample size of 250 participants to estimate robust item parameters (RMSE < .20; Svetina et al., 2013), with an additional 30 cases to deal with dropouts and technical issues. However, due to recruitment problems, we adjusted this goal during the data collection period to meet at least an RMSE < .25, resulting in a targeted sample size of 200 adults. A sample of n = 213 individuals (59 % female) was realized. Participants were 16 to 80 years old (M_{age} = 31.06, SD_{age} = 14.89) with the median at the age of 25, indicating a right-skewed distribution. Most participants reported higher education entrance qualification (47 %) or a university degree (37 %) as the highest educational achievement. Participants were required to be able to read and understand German-language texts and were recruited through flyers and postings on social networks, online advertisements, local newspaper ads, and public marketplaces. Participation was compensated with 30 € and a further 5 € for bringing along other participants.

TPS Test

The developed TPS test is based on Stemmann's (2016) simulations of home automation and appliances. To include at least two devices per type, we focused on adapting devices from three of the original five device types (household appliances, home automation, information and communication technology). To further enhance participants' immersive experience (see Lombard et al., 2009), we added sound effects to the simulations (e.g., mechanical button clicks, and starting noises). To quantify complexity and intransparency, we determined factor scores for both device characteristics from Principal Component Analysis (PCA) based on multiple complexity and intransparency indicators. For example, the complexity indicators included the number of control elements and menu items to reflect the physical properties of the devices (Beckmann & Goode, 2017); the intransparency indicator included variables that represented the labeling of control elements or the number of control elements with multiple uses. The complexity scores from PCA ranged between −1.09 and 2.15; intransparency scores were between −0.67 and 2.03. Further details are described in the Electronic Supplemental Material (ESM; section PCA).

The adapted test includes nine TPS units, each containing a device with four items (one item from the original test and three newly developed items). In line with our definition of TPS, participants had to interact with a device to solve a provided technical problem that required operations beyond routine interactions. When developing new items, particular attention was paid to designing them using different solution principles to allow the items to be locally independent. In the TPS unit "heating control", for example, setting the time and date required a different menu access and use than switching between automatic and manual mode. The items differed according to the number of actions required to solve them.

Working on a TPS unit followed a fixed procedure (ESM; section TPS screenshot). Each unit started by presenting participants with a picture of a device (e.g., digital heating thermostat) and asked about their prior knowledge. Afterwards, participants were instructed to familiarize themselves with the device as much as possible (knowledge acquisition phase, max. 10 minutes). They could independently explore the device and look up functions and features in a manual. The manuals were created based on German norms (DIN EN 82079-1:2021-09; IEC/IEEE, 2021) and included information about device operation, safety issues, and disposal. After the knowledge acquisition phase (i.e., after participants had the opportunity to familiarize themselves with a device but before they completed the test items), participants were informed that they would be given four items and asked to estimate how many items they would solve correctly. This was part of assessing participants' prospective metacognitive judgment (see Dentakos et al., 2019). Finally, participants were presented with the four items requesting them to put the device in a specific target state (knowledge application phase, max. 5 minutes per item, e.g., "Switch to manual mode and set the heating to 18°C.").

Measures

The following section describes the measures used in this study (details in Table ESM 4).

Knowledge Acquisition Completeness

We determined completeness scores for exploration and manual use. Both scores reflect how extensively participants made use of the knowledge acquisition phase and can range between 0 and 1. Exploration completeness was the share of visited device states (e.g., "set day temperature" or "set weekday" for TPS unit "heating control") to the total number of states per TPS unit. Manual use completeness was the share of visited manual pages to the total number of manual pages per TPS unit.

Knowledge Application

The dichotomous item scores of the 36 TPS items were used as the dependent variable. Only if participants fulfilled all steps required to solve a provided technical problem, the item was scored as solved correctly.

Prior Knowledge

Device-specific prior knowledge was assessed via self-reports on one 4-point and two 5-point Likert scales. The items referred to the possession of, knowledge about, and familiarity with the respective device. Higher scores indicated more device-specific prior knowledge. Sum scores were computed across the three items per TPS unit. They showed reliability scores between ω = .67–.81.

Prospective Metacognitive Judgment

We defined three categories based on calibration scores (Silaj et al., 2021) representing high accuracy (reference category), underconfidence, and overconfidence.

Other Measures

Sociodemographic data included gender (male, female, diverse), age, native language (German or other), and educational level (categories from *no graduation* to *academic degree*). The experience of immersion was assessed using three 5-point Likert items (from 1 = *I totally disagree* to 5 = *I totally agree*) with a reliability score of ω = .55. We also asked the participants how they usually familiarize themselves with devices of home automation and appliances.

Design

The participants worked on all TPS units that were applied in a minimally balanced repeated treatment design (Frey et al., 2009). Sociodemographic data was assessed prior to the TPS test, and the experience of immersion after the TPS test. To reduce missing data, each item had to be answered before continuing with the next part.

Analyses

To evaluate the psychometric properties of the TPS test, we conducted analyses based on Item Response Theory (IRT). The one-dimensionality of TPS test items was evaluated using a modified parallel analysis (Drasgow & Lissak, 1983)—a parallel analysis based on the tetrachoric correlations between the dichotomous item scores. To examine local independence for the TPS test items, we examined whether the residual correlations between item pairs exceeded a cutoff value of $r = |0.2|$ (Q3 statistics; Chen & Thissen, 1997). Item infit was considered acceptable within the range of 0.7–1.3 (Wright & Linacre, 1994). To evaluate the impact of device and person-by-device characteristics, we tested generalized linear mixed models (GLMMs). For those, we specified random intercepts for persons, items, and devices. In contrast to the preregistration, we did not include immersion as a control variable in the GLMMs due to the scale's low reliability. An overview of all analyses can be found in Table ESM 5.

Results

Psychometric Properties

The probability of successfully solving the TPS test items ranged from 9.39% to 95.28%. Grouped by device type, Figure 1A shows the item difficulties from a Rasch model. The majority of items were of medium or low difficulty, ranging between −3.86 and 2.87 ($Mdn = -0.85$). The point-biserial correlations of the item scores with the total sum score range between .24 and .56 ($Mdn = .43$). Infit ranged from 0.83 to 1.24, except for item 23 (unit "car infotainment system") which showed an infit of 0.69 and was therefore unsuited for further use. Item difficulty did not seem to systematically differ between device types. Further details are presented in Table ESM.6.

Supporting the assumption of local independence, Q3 item pair statistics showed values above the cutoff in 1.64% of the cases, with the highest residual correlation between items 5 and 7 from the unit "dishwasher ($r = .64$)." A modified parallel analysis showed a non-significant result (2nd eigenvalue observed = 2.49, 2nd eigenvalues averaged across 100 Monte Carlo samples = 2.64, $p = .644$), supporting the expected one-dimensionality (hypothesis H1). Weighted likelihood estimates (WLE) of participants' TPS skills showed a high reliability of .89 ($SD_{WLE} = 1.51$, $Min = -6.11$, $Max = 3.50$). In addition to the TPS skill distribution as a histogram in Figure 1B and the test information function in Figure 1C, we report the conditional reliabilities in Figure 1D which range from 0 to 1 and were derived from the information function (Nicewander, 2018).

We exploratively compared the results from the Rasch model with a 2PL model and a testlet model that considers the unit structure (Table ESM.7). AIC and likelihood ratio test supported the 2PL model, BIC the Rasch model.

Device Characteristics and Person-by-Device Characteristics

Table 1 summarizes the results from the GLMMs. The baseline model indicated that the probability for an average person to correctly solve an average TPS item was 68.6 %. The predictor model included the device and person-by-device characteristics as predictors to determine their effects. Concerning device characteristics, only the degree of intransparency showed the expected negative effect on item solution rates ($B = -0.90$; H2b) while controlling for the minimal number of actions necessary to correctly solve an item ($B = -0.01$). The effect of complexity was not significant ($B = -0.04$; H2a).

Concerning person-by-device characteristics, both completeness of exploration ($B = 0.49$) and manual use ($B = 0.22$) showed the expected positive effects on item solution rates, supporting H3a and H3b. Additionally, and in line with our expectations, device-specific prior knowledge was also positively associated with item solution rates ($B = 0.28$; H3c). Our hypotheses on prospective metacognitive judgment were only partly supported by the GLMM results. While overestimation showed the expected negative association with item solution rates, supporting H3d ($B = -2.15$), underestimation showed an unexpected significant positive effect on item solution rates ($B = 2.13$).

Descriptive results for the variable immersion suggested that the participants' test-taking efforts corresponded with their real-world experiences ($M = 3.59$, $SD = 0.89$) and behavior ($M = 3.63$, $SD = 0.98$; Table ESM.4). The average immersion correlated positively with TPS skill ($r = .21$, $p = .003$). Results of further exploratory analyses regarding the measures of knowledge acquisition and alternative GLMMs can be found in the ESM (section Exploratory analyses).

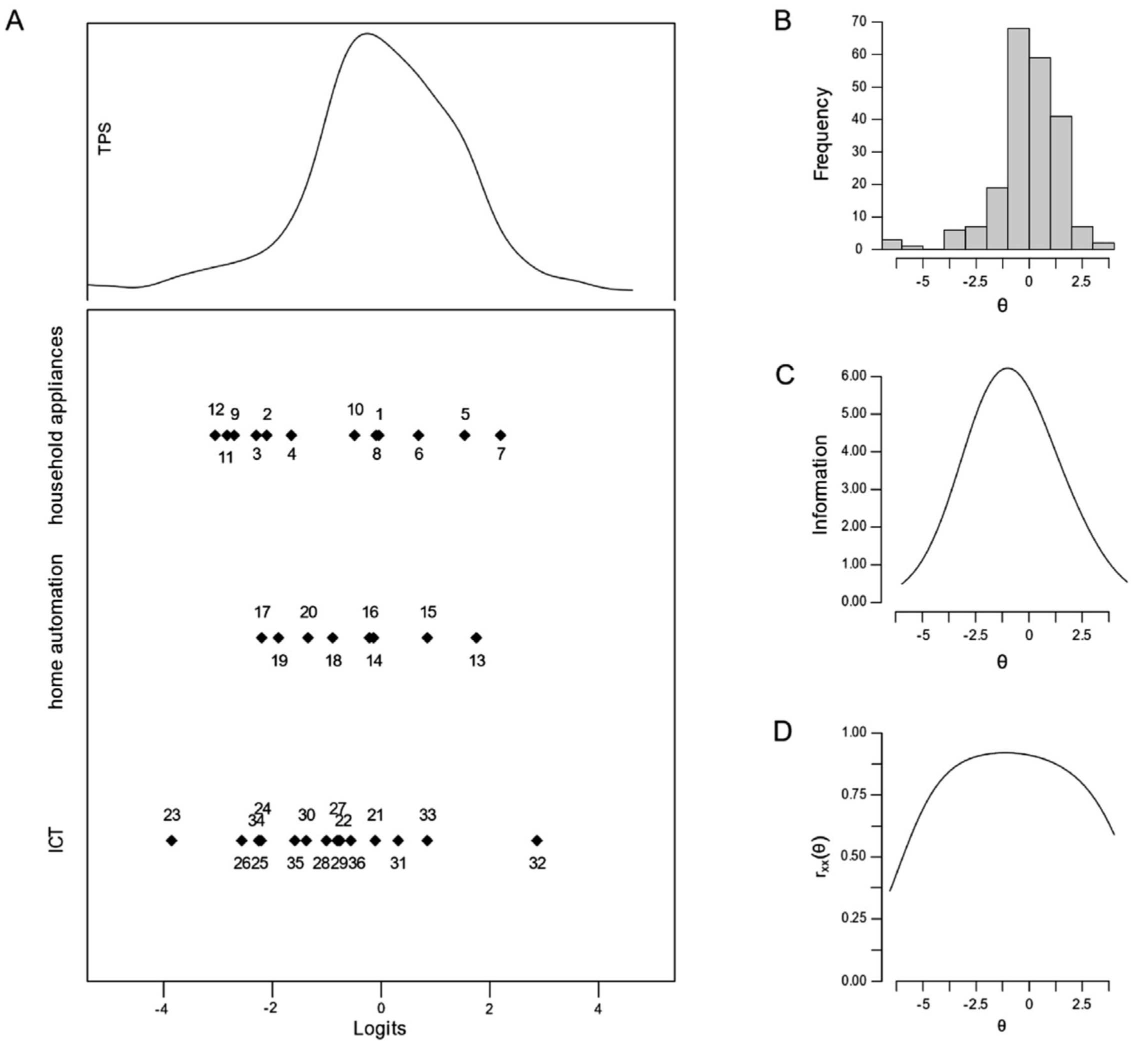

Figure 1. Wright map (A) of the distribution of adults' TPS skill (top of A) mapped on the same scale as the difficulties of the TPS items (bottom of A), distribution of adults' TPS skill (B), Test Information Function (C) and conditional reliabilities (D). Item difficulties in A are clustered according to their device type (y-axis). ICT = information and communication technology. $r_{xx}(\theta)$ = conditional reliabilities.

Discussion

Solving technical problems is a relevant skill in everyday life. In order to investigate TPS skills and behaviors in-depth, we developed an assessment of TPS by reengineering and improving an existing instrument (Stemmann, 2016) in multiple ways. Particular attention was paid to incorporating new items, operating manuals, and measures of device-specific prior knowledge. We analyzed the test's internal structure and effects of device and person-by-device characteristics on item success to validate the interpretation of the resulting TPS score as adults' skill in applying knowledge to solve technical problems.

In line with our expectations, the data showed evidence for the assumed one-dimensional internal structure while providing a highly reliable estimation of individual TPS skills. These results indicate that the TPS items refer to the same skill of solving technical problems. We could also extend previous findings regarding the positive relationship between the completeness of systematic exploration and successful knowledge application (Stemmann, 2016) by demonstrating independent positive effects of manual use and prior knowledge on item success while controlling for immersion experience.

However, our expectations were only partially met regarding the complexity and intransparency characteristics (see Funke, 2001). While a higher degree of intransparency was associated with a lower probability of successful knowledge application, complexity was not. However, the missing effect might be an artifact of our definition of complexity. Following Stemmann (2016), we focused on a device's quantitative extent and used count indicators, such as the

Table 1. Results of the GLMM for effects of device and person-by-device characteristics on item solution probability.

	Baseline model			Predictor model		
Fixed Effects	*B* (*SE*)	*p*	95% CI	*B* (*SE*)	*p*	95% CI
Intercept	0.86 (0.39)	.026*	[0.10, 1.61]	0.89 (0.31)	.004**	[0.29, 1.49]
Device characteristics						
Complexity	–			0.05 (0.24)	.841	[−0.42, 0.52]
Intransparency	–			−0.75 (0.24)	.002**	[−1.23, −0.28]
Person-by-device characteristics						
Completeness exploration	–			0.49 (0.05)	<.001***	[0.40, 0.59]
Completeness manual use	–			0.25 (0.05)	<.001***	[0.15, 0.34]
Prior knowledge	–			0.28 (0.04)	<.001***	[0.21, 0.36]
Metacognitive underestimation	–			2.12 (0.12)	<.001***	[1.88, 2.37]
Metacognitive overestimation	–			−2.13 (0.15)	<.001***	[−2.41, −1.84]
Control variables						
Minimal number of actions	–			−0.01 (0.01)	.453	[−0.03, 0.01]
Random Intercepts						
SD random device intercepts	0.94			– [1]		
SD random item intercepts	1.20			1.26		
SD random person intercepts	1.37			0.95		

Note. n = 213 for the Baseline model. n = 210 for the Predictor model. The reference category of the categorial variable is “accurate metacognitive estimation”. [1] A model including random intercepts for devices revealed a singular fit. Accordingly, the specified random effect structure is overly complex. There is no systematic effect coming from the device, indicating that the model’s predictors explained the variance of device level completely. Thus, we removed the random intercept for the device. Results did not differ between models including vs. not including random intercepts for devices. *p < .05; **p < .01; ***p < .001.

number of control elements or states to create complexity scores (Table ESM.2). Other approaches also consider situational characteristics such as instructions to evaluate a task’s complexity (Beckmann & Goode, 2017). Another explanation relates to the level of the salience of complexity and intransparency for participants. Adults can quickly estimate a device’s complexity when facing a simulation, for example, by exploring the functions and menus, reading or skimming the manual, or simply inferring complexity from the number of control elements. Consequently, they might adapt their effort. Due to their nature, and contrary to complexity, intransparent aspects are challenging to be perceived. Accordingly, participants might have to be initially aware that functions and menu items exist that are not obvious to detect. If they do not encounter intransparent aspects of a device, participants might have carried this lack of operational knowledge over to the knowledge application phase, which would not be the case for complexity.

Post-hoc analyses (Table ESM.8-9) further revealed that participants spent more time in the knowledge acquisition phase in more complex devices. This effect is even stronger for intransparency. Additionally, complexity was highly negatively correlated with the completeness of exploration but positively with manual use, despite not having shown an effect on item success in the GLMM analysis. Intransparency did not affect the completeness of exploration or manual use. Although these results require further investigation, they at least suggest that the successful application of operational knowledge in TPS is due to the active acquisition of operational knowledge (e.g., see Greiff et al., 2012) and the activation of prior knowledge (see Klahr & Dunbar, 1988).

Regarding adults’ prospective metacognitive judgments as the last person-by-device characteristic regarded in this study, expectations were partially met. Dentakos et al. (2019) suggested that prospective metacognitive judgments can affect how cognitive resources are allocated and that both overestimation and underestimation should have negative consequences for problem solving. However, our results demonstrate only the negative effect of metacognitive overestimation. Metacognitive underestimation, on the other hand, was associated with more successful knowledge application. This might be an instance of the Dunning-Kruger effect, meaning that people with low abilities tend to overestimate their abilities, and people with high abilities tend to underestimate them (e.g., Magnus & Peresetsky, 2022). Another explanation would be that metacognitive underestimation may lead to more effort in problem solving thus improving test performance.

Limitations

First, computer simulations of home automation and appliances are not likely to provide adults with the same experiences as real devices (e.g., haptic sensations, and device feedback). However, despite low reliability, participants reported rather high immersion. Second, the TPS environments might have affected the way adults would typically use operating manuals as it was not possible to read manuals simultaneously with operating a device. Instead, the manual needed to be closed to interact with the simulation. Third, there is a potential overlap of complexity (device characteristic) and the number of actions required to solve an item (item characteristic) that cannot be ruled out completely. However, the correlation of the item characteristic with a disaggregated complexity score speaks at least against a dependency between both variables (r = .09). Fourth, regarding the PCA for intransparency scores, the scree plot was ambiguous. We still decided to integrate all indicators into a single factor score for a better representation of intransparency. Fifth, regarding metacognitive accuracy, we only assessed prospective but not retrospective confidence judgments. Additionally, we categorized the determined calibration score (see Silaj et al., 2021) to be able to interpret the results from the conducted GLMMs. However, this treatment might have resulted in information loss. Sixth, we elaborated on the relationship of TPS with other cognitive constructs, such as intelligence, CPS, and ICT skills, but did not investigate these variables. Future research will need to provide valid evidence about the assumed nomological net. Finally, the composition of our ad-hoc sample provides limitations, as the age distribution was right-skewed and most participants reported a high level of education. Accordingly, our results may not be generalized to older adults and adults with lower levels of education.

Conclusion

Overall, our results indicate a successful test development in multiple ways. First, the TPS scale was one-dimensional across nine TPS units and a total of 36 items. As item difficulties did not seem to differ between device types systematically, economic test application is possible (i.e., reducing the test time by using fewer units). Second, we extended previous findings on the relationship between knowledge acquisition and knowledge application (Stemmann, 2016) by considering the effects of manual use and prior knowledge as predictors. Third, we found unique positive effects for completeness of manual use and device-specific prior knowledge while considering variations on person, device, and item level. Despite some unexpected findings (e.g., the missing effect of complexity) and multiple limitations, there is valid evidence concerning the internal test structure, response processes, and relations to other variables (see AERA, APA, & NCME, 2014) that speak in favor of interpreting the TPS test scores as adults' skill to purposefully apply operational knowledge when solving a technical problem.

References

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitzschke, A., & Funke, J. (2012). Dynamische Problemlösekompetenz: Ein bedeutsamer Prädiktor von Problemlöseleistungen in technischen Anforderungskontexten? [Dynamic problem solving – an important predictor of problem solving performance in technical domains?]. *Zeitschrift für Erziehungswissenschaft, 15*(2), 363–391. https://doi.org/10.1007/s11618-012-0277-9

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Beckmann, J. F., & Goode, N. (2017). Missing the wood for the wrong trees: On the difficulty of defining the complexity of complex problem solving scenarios. *Journal of Intelligence, 5*(2), 15. https://doi.org/10.3390/jintelligence5020015

Binkley, M., Erstad, O., Herman, J., Raizen, S., Riple, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer Netherlands. https://doi.org/10.1007/978–94-007–2324-5_2

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education, 53*(4), 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265. https://doi.org/10.2307/1165285

Dentakos, S., Saoud, W., Ackerman, R., & Toplak, M. E. (2019). Does domain matter? Monitoring accuracy across domains. *Metacognition and Learning, 14*(3), 413–436. https://doi.org/10.1007/s11409-019-09198-4

Dörner, D. (1976). *Problemlösen als Informationsverarbeitung* [Problem solving as information processing] (1st ed). Kohlhammer.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363–373. https://doi.org/10.1037/0021-9010.68.3.363

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning, 7*(1), 69–89. https://doi.org/10.1080/13546780042000046

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*(3), 189–213. https://doi.org/10.1177/0146621612439620

IEC/IEEE. (2021). *Preparation of information for use (instructions for use) of products—Part 1: Principles and general requirements (IEC/IEEE 82079–1:2019); German version EN IEC/IEEE 82079–1:2020* (Issues 82079–1). Beuth.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48. https://doi.org/10.1207/s15516709cog1201_1

Kretzschmar, A., Hacatrjana, L., & Rascevska, M. (2017). Re-evaluating the psychometric properties of MicroFIN: A multidimensional measurement of complex problem solving or a unidimensional reasoning test? *Psychological Test and Assessment Modeling, 59*(2), Article 157182. https://doi.org/10.5167/uzh-185323

Lindqvist, E., PerssonVasiliou, A., Hwang, A. S., Mihailidis, A., Astelle, A., Sixsmith, A., & Nygård, L. (2018). The contrasting role of technology as both supportive and hindering in the everyday lives of people with mild cognitive deficits: A focus group study. *BMC Geriatrics, 18*(1), 185. https://doi.org/10.1186/s12877-018-0879-z

Lombard, M., Ditton, T. B., & Weinstein, L. (2009). Measuring presence: The temple presence inventory. *Proceedings of the 12th Annual International Workshop on Presence*, 1–15.

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action–effective strategic behaviors while solving complex problems. *Intelligence, 64*, 98–112. https://doi.org/10.1016/j.intell.2017.08.002

Magnus, J. R., & Peresetsky, A. A. (2022). A statistical explanation of the Dunning–Kruger Effect. *Frontiers in Psychology, 13*, Article 840180. https://doi.org/10.3389/fpsyg.2022.840180

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Mihalca, L., & Mengelkamp, C. (2020). Effects of induced levels of prior knowledge on monitoring accuracy and performance when learning from self-regulated problem solving. *Journal of Educational Psychology, 112*(4), 795–810. https://doi.org/10.1037/edu0000389

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods, 23*(2), 351–362. https://doi.org/10.1037/met0000132

Silaj, K. M., Schwartz, S. T., Siegel, A. L. M., & Castel, A. D. (2021). Test anxiety and metacognitive performance in the classroom. *Educational Psychology Review, 33*(4), 1809–1834. https://doi.org/10.1007/s10648-021-09598-6

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence, 72*, 1–12. https://doi.org/10.1016/j.intell.2018.11.003

Stemmann, J. (2016). *Technische Problemlösekompetenz im Alltag – Theoretische Entwicklung und empirische Prüfung des Kompetenzkonstruktes*. [Technical problem solving skills in everyday life: Theoretical development and empirical examination of the competence construct 'Problem Solving in dealing with technical devices']. University Duisburg-Essen.

Süß, H.-M., & Kretzschmar, A. (2018). Impact of cognitive abilities and prior knowledge on complex problem solving performance – Empirical results and a plea for ecologically valid microworlds. *Frontiers in Psychology, 9*, Article 626. https://doi.org/10.3389/fpsyg.2018.00626

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., Gorin, J. S., Fay, D., & Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling, 55*(4), 335–360.

Tamas, R., O'Brien, W., & Quintero, M. S. (2021). Residential thermostat usability: Comparing manual, programmable, and smart devices. *Building and Environment, 203*, Article 108104. https://doi.org/10.1016/j.buildenv.2021.108104

Wirth, J. (2004). *Selbstregulation von Lernprozessen* [Self-regulation of learning processes]. Waxmann.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370–371.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving – more than reasoning? *Intelligence, 40*(1), 1–14. https://doi.org/10.1016/j.intell.2011.11.003

History
Received June 30, 2023
Revision received April 19, 2024
Accepted May 8, 2024
Published online August 8, 2024
EJPA Section / Category Educational Psychology

Acknowledgments

We thank all student assistants as well as all participants for their valuable contributions to this study.

Publication Ethics

All procedures in studies involving human participants were performed in accordance with the ethical standards of the local Ethics Committee (votum number DIPF_EK_2022_19).

Authorship

Marcus Schrickel, writing – original draft, formal analysis, software; Jennifer Stemmann, writing – review & editing, conceptualization, software; Frank Goldhammer, writing – review & editing, methodology; Carolin Hahnel, writing – review & editing, formal analysis, conceptualization, supervision.
All authors approved the final version of the article.

Open Science

We report all measures in the study and all analyses including all tested models. If we use inferential tests, we report exact *p* values.
Open Data: The data and the instrument are available via the German Network of Educational Research Data under https://doi.org/10.7477/1136:401:62 (Hahnel et al. 2024).
Open Materials: The information needed to reproduce this study is available at https://osf.io/ux6y4 (Schrickel et al., 2024).
Open Analytic Code: I confirm that all the scripts, code, and outputs needed to reproduce the results are available at https://osf.io/ux6y4 (Schrickel et al., 2024).
Supplemental Materials: Supplemental materials are available at https://osf.io/ux6y4 (Schrickel et al., 2024).
Preregistration of Studies and Analysis Plans: This study was preregistered with an analysis plan on the Open Science Framework and can be found at https://doi.org/10.17605/OSF.IO/HC7ZX (Schrickel et al., 2023).

Funding

The research was funded by the German Research Foundation (DFG), grant number 456978965. Open access publication enabled by Ruhr University Bochum.

ORCID
Marcus Schrickel
https://orcid.org/0000-0002-9424-8220
Jennifer Stemmann
https://orcid.org/0000-0002-9105-579X
Frank Goldhammer
https://orcid.org/0000-0003-0289-9534
Carolin Hahnel
https://orcid.org/0000-0003-2394-3944

Marcus Schrickel
Faculty of Psychology
Department of Psychological Assessment and Testing
Ruhr University Bochum
Universitätsstraße 150
44801 Bochum
Germany
marcus.schrickel@rub.de