Take your time: Invariance of time-on-task in problem solving tasks across expertise levels

Matthias Stadler[1], Anika Radkowitsch[1], Ralf Schmidmaier[2], Martin Fischer[2] & Frank Fischer[1]

[1] Fakultät für Psychologie und Pädagogik Ludwig-Maximilians-Universität München, Munich, Germany

[2] Klinikum der Universität, Ludwig-Maximilians-Universität München, Munich, Germany

Abstract:

Computer-based tasks provide a vast amount of behavioral data that can be analyzed in addition to the indicators of final performance. One of the most commonly investigated indicators is time-on-task (ToT), which is understood as the time from task onset to task completion. Studies often assume a unidimensional measurement model with one latent ToT variable that is sufficient to capture all response time covariance across items. However, behavioral indicators such as ToT are seldom submitted to the same psychometric rigor as more traditional indicators. In this brief report, we provide first results on the invariance of ToT in problem-solving tasks across different levels of expertise. A total of 98 medical students and physicians participated in the study quasi-experimentally grouped into three conditions (low, intermediate, and high) based on their prior knowledge. All participants solved five medical diagnostic problem-solving tasks in a simulation-based learning environment. While the overall ToT seems to decrease with level of expertise, the general pattern across tasks seems to be similar for all three groups. The results indicate strong measurement invariance of ToT across different levels of expertise and support interpreting group differences on a latent ToT factor.

Keywords:

Time-on-task, Behavior, Invariance, Expertise, Assessment

Take your time: Invariance of time-on-task in problem solving tasks across expertise levels

Computer-based tasks provide a vast amount of behavioral data that can be analyzed in addition to the indicators of final performance (Bunderson, Inouye, & Olsen, 1988). As early as 1953, Ebel noted the "possible usefulness of response time data in selecting the most efficient items, and in otherwise probing the nature and functions of the test items". Such response time data or time-on-task (ToT) is still among the most widely used indicators of task behavior. ToT is understood as the time from task onset to task completion. Thus, if the task was completed in order, it reflects the time taken to become familiar with the task, to process the materials provided to solve the task, to think about the solution, and to give a response (Goldhammer et al., 2014). Analyzing individual differences in ToT may thus allow researchers to make inferences from overt behavior about the latent cognitive processes involved in solving tasks (Goldhammer & Zehner, 2017). In interpreting ToT, studies often assume a unidimensional measurement model with one latent ToT variable that is sufficient to capture all response time covariance across items (e.g., Molenaar, 2014; Rudolph, Greiff, Strobel, & Preckel, 2018). However, even though behavioral indicators such as ToT are used as additional evidence for process oriented constructs (Klein Entink, Fox, & van der Linden, 2009), they are not submitted to the same psychometric rigor as more traditional indicators. Questions such as the scalability or individual stability of task behavior remain to be answered. When measurements are not scaled equivalently, analyses of individual differences may not only reflect the phenomena of interest (e.g., construct-level relationships or change) but also systematic variation in measurement. It is thus essential to determine whether task behavior, just like other measures, produce comparable measurements for all individuals within the population under study (Bauer, 2017; Greiff & Scherer, 2018).

Expertise in time-on-task

In this brief report, we provide first results on the invariance of ToT in problem-solving tasks across different levels of expertise as an empirical example on how psychometric standards can be applied to process data. In skill assessments, the relation between ToT and task result can be conceived of in two ways. On the one hand, taking more time to work on a task may be positively related to the result as the task is completed more thoroughly. On the other hand, the relation may be negative if working faster and more fluently reflects a higher skill level (Goldhammer et al., 2014). Previous studies demonstrated that problem-solvers level of expertise moderates the ToT as a function of the tasks' difficulty. Spending more time was associated with higher probability of solving a problem correctly when dealing with difficult problems. Problem-solvers could benefit more from spending more ToT when their individual level of expertise was low than when it was already high. These results seem to indicate that poor problem-solvers may be able to compensate for their lack of expertise by spending more effort on a solving a particular problem. This effect was especially strong when the problem was difficult (Goldhammer et al., 2014; Scherer, Greiff, & Hautamäki, 2015).

Such strong sources of systematic inter-individual variance raise the question,

though, whether the same underlying latent construct is causing individual differences in ToT or whether differences in ToT reflect different constructs depending on level of expertise. If the meaning of ToT changes from indicating motivation to indicating skill with increasing level of expertise, a latent ToT factor may not be comparable across levels of expertise. We investigated this research question by estimating measurement invariance of a latent ToT factor for five knowledge rich problem-solving tasks across three different levels of expertise.

Methods

Sample and Design

A total of 98 medical students and physicians participated in the present study of whom 61 were female. All participants were quasi-experimentally grouped into three conditions based on their prior knowledge. The *low prior knowledge* group consisted of 45 medical students in their 1st and 2nd clinical year of medical school (M = 6.4, SD = 0.7). The *intermediate prior knowledge* group consisted of 27 medical students in their 3rd and final clinical year of medical school (M = 11.1, SD = 1.7). The *expert* group consisted of 26 physicians with a specialization in internal medicine with a minimum of three years of working experience (M = 14.5, SD = 10.8).

Procedure

After participants had given their informed consent, all participants first completed a knowledge test assessing their medical knowledge. This test was used as a manipulation check for the quasi-experimental grouping of participants. Afterwards, all participants solved five medical diagnostic problems that were implemented in a simulation-based learning environment. Participants completed both tasks (the medical knowledge test and the medical diagnostic problem-solving tasks) computer based. The instruction was standardized for all participants. Afterwards, all participants were thanked, debriefed, and dismissed.

Measures

Medical knowledge

We assessed the participants' medical knowledge by using eight so-called key feature cases (Fischer, Kopp, Holzer, Ruderich, & Jünger, 2005). Each key feature case consists of a short case vignette describing key features of a medical problem and three questions with respect to the most plausible diagnoses as well as the most important diagnostic and treatment steps. For each of the 24 items, participants could receive 0 or 1 point. The mean medical knowledge was calculated for each participants (range between 0 and 1).

Time on Task

All participants solved five medical diagnostic problem-solving tasks in a simulation-based learning environment. The task was to diagnose five consecutive medical patient cases by naming and justifying a main diagnosis and by suggesting further differential diagnoses as well as further treatment steps. For each patient case, the participants first received a detailed electronic health record, which contained information about findings, symptoms, and laboratory values. Afterwards, the participants needed to interact with a simulated radiol-

ogist in order to generate further evidence. The simulated radiologist responded to the participants request only if the request was properly justified. Finally, the participants were asked to document the results of their diagnostic reasoning process in the electronic health record as described above. The ToT for each patient case was generated from log files.

Statistical analyses

All analyses were done using MPlus 7.0 (Muthén & Muthén, 1998-2012) and R 3.5.2 (R Core Team, 2018). To assess whether the quasi-experimentally grouped prior-knowledge conditions differed with respect to their medical knowledge, we conducted a one-way ANOVA with medical knowledge as dependent and condition as independent variable. Bonferroni post-hoc tests were conducted to analyze differences between groups.

The ToT for all five tasks were defined to load onto one latent factor. For the configural model, factor loadings were estimated freely for the three levels of expertise. To test for metric invariance, factor loadings were constrained to be equal across the three groups. For scalar invariance, loadings and intercepts were constrained to be equal (Cheung & Rensvold, 2002). To evaluate, whether the constraints lead to significant reduction in model fit, we compared chi-squared values and McDonalds non-centrality-index (McDonald, 1989). Following the suggestions by Cheung and Rensvold (2002), we defined ΔNFI > .02 to indicate non-invariance across models. We refrained from comparing models based on their adjusted root mean squared errors (RMSEA) due to the relatively low sample size and few degrees of freedom (Kenny, Kaniskan, & McCoach, 2015). The data and scripts for all analyses can be found on the open science framework repository (blinded for review).

Results

Preliminary Results

The prior knowledge groups differ descriptively with respect to their medical prior knowledge (see Table 1) with participants of the Expert group showing higher medical knowledge compared to the intermediate and the low prior knowledge groups. The ANOVA shows that groups differ significantly ($F(2,98) = 38.89$, $p < .001$, $eta^2 = 0.44$). Post-hoc analyses show that the low prior knowledge group differs significantly from the expert group (95% CI [-0.28; -0.15]) and the intermediate group (95% CI [-0.23; -0.10]). However, the intermediate and expert group did not differ significantly (95% CI [-0.12; 0.02]).

Main Results

Figure 1 displays the average ToT in minutes across the five tasks for each level of expertise. As can be seen, ToT varies across tasks. While the average ToT decreases with level of expertise ($F[2,95] = 9.28$, $p < .001$), the general pattern across tasks seems to be similar for all three groups. Regardless of level of expertise, average ToT is highest for Tasks 3 and 5 and lowest for Task 4. This pattern also reflects in the factor loadings for the configural model. Table 1 shows the factor loadings for the configural model with all tasks loading on one common factor and loadings estimated freely.

The configural model showed a good fit to the data ($X^2[12] = 16.2$, $p = .180$, NCI = .978).

Constraining factor loadings to equal for the metric model ($X^2[20] = 20.5$, $p = .425$, NCI = .997) did not change the fit significantly ($\Delta X^2[8] = 5.2$, $p = .741$, ΔNFI = .019). Finally, loadings and intercepts were constrained to be equal for the scalar model ($X^2[28] = 34.1$, $p = .200$, NCI = .969). This model also did not fit the data significantly worse than the configural model ($\Delta X^2[16] = 18.2$, $p = .314$, ΔNFI = .007).

Table 1 Means and standard deviations of the medical knowledge per group.

Level of expertise	Mean	SD
Low	.45	.11
Intermediate	.61	.10
Expert	.66	.12

Table 2 Standardized factor loadings of the configural model

Level of expertise	Task 1	Task 2	Task 3	Task 4	Task 5
Low	.65	.80	.56	.41	.25
Intermediate	.64	.90	.51	.46	.30
Expert	.63	.91	.55	.59	.20

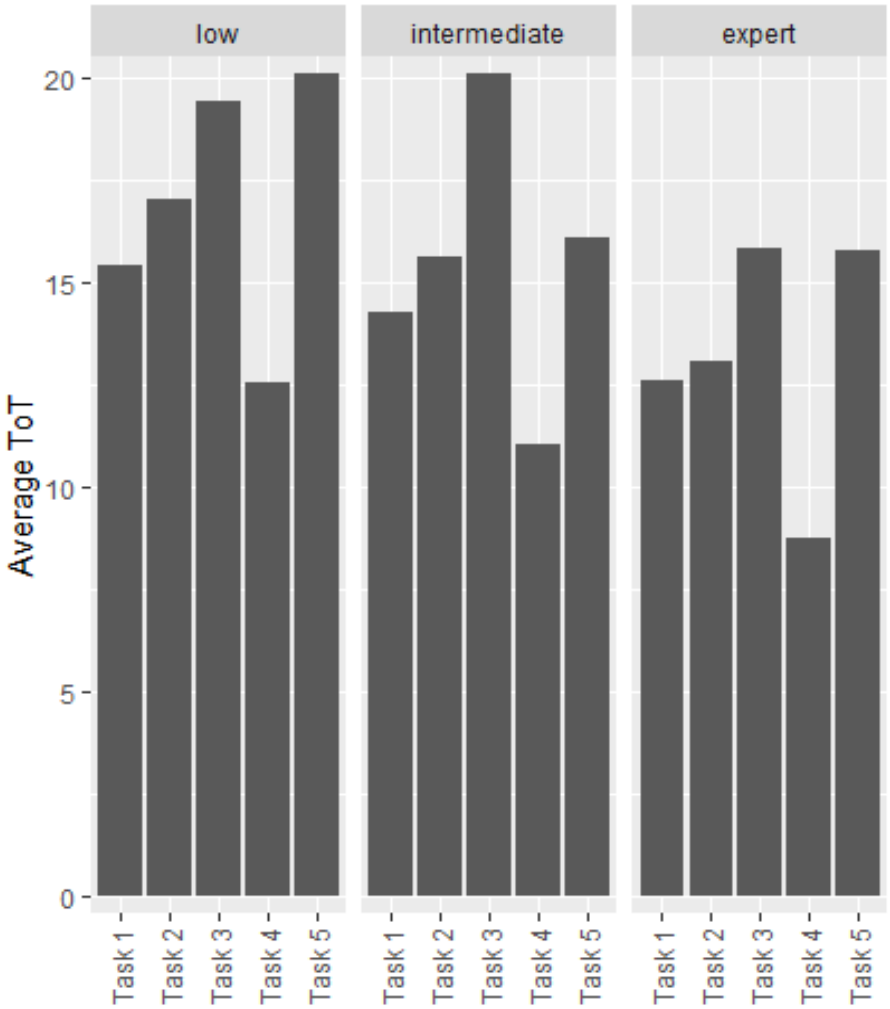

Figure 1 Distribution of average time-on-task (ToT) in minutes across the tasks for each level of expertise

Discussion

The aim of this study was to provide a brief example of applying psychometric standards to indicators of task behavior. Specifically, we investigated whether a latent factor based on ToT scores of five problem-solving tasks was invariant across different levels of expertise. We found that, as expected, ToT decreased with increasing level of expertise on all five tasks. However, the tasks' loadings on a common factor did not vary significantly across levels of expertise. This implies that changes in difficulty across the tasks caused equal changes in ToT regardless of expertise. These findings support the use and interpretation of a latent ToT-Factor and allow analyzing differences in ToT between different levels of expertise on a latent level. However, the findings are seemingly at odds with previous findings on the relation between ToT and expertise that would have predicted problem-solvers with different levels of expertise to change their ToT differently depending on task difficulty (Goldhammer et al., 2014; Greiff & Scherer, 2018). As a potential solution to this seeming contradiction, Naumann (2019, April 1st) investigated whether student dispositions other than the targeted skill might impact students' time on task behavior. Following their argument, individual differences in motivation rather than differences in skill may moderate the relation between task difficulty and ToT. Given the game-like character of the problem-solving tasks used in this study and the tasks' relevance to the participants' field of work, motivation should have been high for all participants. This may explain why the pattern of average ToT across tasks and, in that, the relation between task difficulty and ToT was invariant across levels of expertise in our study but not others.

A clear limitation of our study is the quasi-experimental attribution of participants to the levels of expertise based on years of experience (Schmidt & Boshuizen, 1993). Although informative and plausible, this approach is a relatively weak one, because it does not allow for active manipulation of variables of interest. Previous studies have, however, demonstrated the validity of assuming fundamental differences in the expertise of medical students and practitioners based on their experience (Verkoeijen, Rikers, Schmidt, van de Wiel, & Kooman, 2004). The results of our study further support the chosen distinction with continuous decreases in ToT across levels of expertise. In addition, the results of this knowledge test show differences between the three levels of expertise in the expected directions. Nonetheless, future studies may want to corroborate our findings applying an experimental manipulation of expertise (Schmidt & Boshuizen, 1993).

More importantly, the relatively low number of participants limited the statistical power of our invariance analyses. Especially the chi-square difference test is known to be sensitive to sample size and may have underestimated the differences in model fit between the configural, metric and scalar models (Cheung & Rensvold, 2002). Mc Donald's NCI however in less sensible to sample size and provides a more reliable estimate of invariance in our study (McDonald, 1989). The very low complexity of our CFAs with at least medium factor loadings however should result in trustworthy results despite the small samples (Wolf, Harrington, Clark, & Miller, 2013). We believe that the main value of our paper lies not in the empirical example but rather in being a first step in instigating a discussion on the psychometric quality of process data.

Future studies should replicate our findings with a larger sample, though.

In conclusion, this brief report provides an example of applying psychometric standards to indicators of task behavior. As with all studies of invariance, we cannot say whether our findings can be generalized to other problem-solving tasks or even other indicators of behavior than ToT. Therefore, we hope that the study will instigate more research on the psychometric quality of behavioral indicators to fully gauge the validity of these extremely promising data.

References

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. *ETS Research Report Series*, *1988*(1), i-148.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Ebel, R. L. (1953). The Use of Item Response Time Measurements in the Construction of Educational Achievement Tests. *Educational and Psychological Measurement*, *13*(3), 391–401. https://doi.org/10.1177/001316445301300303

Fischer, M. R., Kopp, V., Holzer, M., Ruderich, F., & Jünger, J. (2005). A modified electronic key feature examination for undergraduate medical students: Validation threats and opportunities. *Medical Teacher*, *27*(5), 450–455. https://doi.org/10.1080/01421590500078471

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716

Goldhammer, F., & Zehner, F. (2017). What to Make Of and How to Interpret Process Data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3-4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

Greiff, S., & Scherer, R. (2018). Still Comparing Apples With Oranges? *European Journal of Psychological Assessment*, *34*(3), 141–144. https://doi.org/10.1027/1015-5759/a000487

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*(1), 97–103. https://doi.org/10.1007/BF01908590

Molenaar, I. (2014). Advances in temporal analysis in learning and instruction. *Frontline Learning Research*, *6*, 15–24. https://doi.org/10.14786/flr.v2i4.118

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Naumann, J. (2019, April 1st). *The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment.* Retrieved from https://www.researchgate.net/publication/330534839_The_skilled_the_knowledgeable_and_the_motivated_Investigating_the_strategic_allocation_of_time_on_task_in_a_computer-based_assessment

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Rudolph, J., Greiff, S., Strobel, A., & Preckel, F. (2018). Understanding the link between need for cognition and complex problem solving. *Contemporary Educational Psychology*, *55*, 53–62. https://doi.org/10.1016/j.cedpsych.2018.08.001

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the Relation between Time on Task and Ability in Complex Problem Solving. *Intelligence*, *48*, 37–50. https://doi.org/10.1016/j.intell.2014.10.003

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On acquiring expertise in medicine. *Educational Psychology Review*, *5*(3), 205–221. https://doi.org/10.1007/BF01323044

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., Schmidt, H. G., van de Wiel, M. W. J., & Kooman, J. P. (2004). Case representation by medical experts, intermediates and novices for laboratory data presented with or without a clinical context. *Medical Education*, *38*(6), 617–627. https://doi.org/10.1046/j.1365-2923.2004.01797.x

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

This research was supported by a project funded by the DFG (for2385; COSIMA; Teilprojekte 6 und M).

Corresponding author:
Matthias Stadler, PhD
Leopoldstr. 13
80802 München
Germany
Matthias.Stadler@lmu.de