Goldhammer, Frank; Naumann, Johannes; Stelter, Annette; Tóth, Krisztina; Rölke, Heiko; Klieme, Eckhard

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

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
The Journal of educational psychology 106 (2014) 3, S. 608-626

Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:
urn:nbn:de:0111-pedocs-179679
10.25656/01:17967

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-179679
https://doi.org/10.25656/01:17967

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

Frank Goldhammer[ab], Johannes Naumann[ab], Annette Stelter[a], Krisztina Tóth[a], Heiko Rölke[a], Eckhard Klieme[ab]

[a]German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany
[b]Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany

Abstract

Computer-based assessment can provide new insights into behavioral processes of task completion that cannot be uncovered by paper-based instruments. Time presents a major characteristic of the task completion process. Psychologically, time on task has 2 different interpretations, suggesting opposing associations with task outcome: Spending more time may be positively related to the outcome as the task is completed more carefully. However, the relation may be negative if working more fluently, and thus faster, reflects higher skill level. Using a dual processing theory framework, the present study argues that the validity of each assumption is dependent on the relative degree of controlled versus routine cognitive processing required by a task, as well as a person's acquired skill. A total of 1,020 persons ages 16 to 65 years participated in the German field test of the Programme for the International Assessment of Adult Competencies. Test takers completed computer-based reading and problem solving tasks. As revealed by linear mixed models, in problem solving, which required controlled processing, the time on task effect was positive and increased with task difficulty. In reading tasks, which required more routine processing, the time on task effect was negative and the more negative, the easier a task was. In problem solving, the positive time on task effect decreased with increasing skill level. In reading, the negative time on task effect increased with increasing skill level. These heterogeneous effects suggest that time on task has no uniform interpretation but is a function of task difficulty and individual skill.

Keywords: computer-based assessment, time on task, automatic and controlled processing, reading literacy, problem solving

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG), awarded to Frank Goldhammer, Johannes Naumann, and Heiko Rölke (GO 1979/1-1). We are grateful to Beatrice Rammstedt and her group at GESIS (http://gesis.org), as well as the Federal Ministry for Education and Research (BMBF) for making the data available for this study. Correspondence concerning this article should be addressed to Frank Goldhammer, German Institute for International Educational Research (DIPF), Rostocker Straße 6, 60323 Frankfurt/Main, Germany. E-mail: goldhammer@dipf.de

There are two fundamental observations on human performance: the result obtained on a task and the time taken (e.g., Ebel, 1953). In educational assessment, the focus is mainly on the task outcome; behavioral processes that led to the result are usually not considered. One reason may be that traditional assessments are paper-based and, hence, are not suitable for collecting behavioral process data at the task level (cf. Scheuermann & Björnsson, 2009). However, computer-based assessment— besides other advantages, such as increased construct validity (e.g., Sireci & Zenisky, 2006) or improved test design (e.g., van der Linden, 2005) - can provide further insights into the task completion process. This is because in computer-based assessment, log file data can be recorded by the assessment system that allows the researcher to derive theoretically meaningful descriptors of the task completion process. The present study draws on log file data from an international computer-based large-scale assessment to address the question of how time on task is related to the task outcome. As shown in the following, by analyzing the relation of task performance to the time test takers spent on task, we were able to obtain new insights into how the interaction of task and person characteristics determines the way of cognitive processing. For instance, this can contribute to the validation of the assessment, if time on task can be related to the task response in a theoretically sound way.

Time on task is an important characteristic of the solution process indicating the duration of perceptual, cognitive, and psychomotorical activities. From a measurement point of view, the usefulness of time on task and the task outcome, respectively, depend on the tasks' difficulty. In easy tasks assessing basic skills, individual differences will mainly occur in response latencies, whereas accuracy will be consistently high. Following this logic, a number of assessment tools that address constructs like naming speed (e.g., Nicolson & Fawcett, 1994), visual word recognition (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), or number naming speed (e.g., Krajewski & Schneider, 2009) make use of time on task. In contrast, in more difficult tasks the accuracy of a result is of interest, for example, in assessments of reading comprehension (e.g., van den Broek, & Espin, 2012) or problem solving (e.g., Greiff, Wüstenberg, et al., 2013; Klieme, 2004; Mayer, 1994; Wirth & Klieme, 2003). In these skill assessments, time on task usually is not taken into account. Nevertheless, both the task result and time on task constitute task performance regardless of the task's difficulty.

In skill assessments, the relation between time on task and task result (accuracy) 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. The present study addresses these contradictory predictions and aims at clarifying the conditions of their validity by jointly analyzing task success and time on task data from the computer-based Programme for the International Assessment of Adult Competencies (PIAAC; cf. OECD, 2013; Schleicher, 2008). Thus, we take advantage of the fact that computer-based assessment renders data available on a large scale that was previously available only through small-scale experimenting (i.e., time on task). Data such as time spent on individual tasks can serve to answer basic research questions (such as clarifying the relation of time on task and task result in different domains). Furthermore, the data can enhance educational assessment. For instance, construct validation can be supported by testing whether behavioral process indicators are related to task outcomes as expected from theory.

Time on Task

Time on task 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.[1] In tasks requiring the participant to interact with the stimulus through multiple steps, time on task can be further split into components, for instance, reflecting the time taken to process a single page from a multipage stimulus. To model time on task, two different approaches have been suggested (cf. van der Linden, 2007, 2009). First, time is considered an indicator of a (latent) construct, for example, reading speed (Carver, 1992) or reasoning speed (Goldhammer & Klein Entink, 2011). Here, response and time data are modeled using separate measurement models. Second, within an explanatory item response model, time is used as a predictor to explain differences in task success (cf. Roskam, 1997). In the present study, this second approach is used to investigate the relation between time on task and task success. Task success (dependent variable) can be perceived as a function of time on task (independent variable) because the individual is able to control time spent on completing a task to some extent, which in turn may affect the probability of attaining the correct result (cf. van der Linden, 2009).

Relation of Time on Task to Task Success

When investigating the relation between time on task and task success, the well-known speed–accuracy tradeoff, which is usually investigated in experimental research (cf. Luce, 1986), has to be taken into account. Tradeoff means that for a given person working on a particular task, accuracy will decrease as the person works faster. The positive relation between time on task and task success, as predicted by the speed–accuracy tradeoff, is a within-person phenomenon that can be expected for any task (e.g., Wickelgren, 1977). However, when switching from the within-person level to a population, the relation between time on task and task success might be completely different, for instance, a negative or no relation, although *within* each person, the speed–accuracy compromise remains as the positive relation between time on task and task success (cf. van der Linden, 2007). Consequently, at the population level, findings on the relation of time on task with task success may be heterogeneous. One line of research modeling time on task as an indicator of speed provides speed–skill or speed–ability correlations of different directions and strengths across domains. For example, for reasoning, positive correlations between skill (measured through task success) and slowness (measured through time on task) were found (e.g., Goldhammer & Klein Entink, 2011; Klein

[1] Depending on what is considered to be a task, there may be alternative definitions of time on task. For instance, in this special section, Kupiainen, Vainikainen, Marjanen, and Hautamäki (2014) use the term time on task to refer to the time needed to complete a test in a learning to learn assessment, whereas response time is considered to represent the time needed to respond to a single question or problem (which is comparable to our notion of time on task).

Entink, Fox, & van der Linden, 2009). For arithmetic zero correlations (van der Linden, Scrams, & Schnipke, 1999) were obtained, whereas for basic skills to operate a computer's graphical user interface, a negative relation was demonstrated (Goldhammer, Naumann, & Keßel, 2013), as was for basic reading tasks such as phonological comparison and lexical decision (Richter, Isberner, Naumann, & Kutzner, 2012).

These results suggest that the time on task effect might be moderated by domain and task difficulty. A comparison of tasks across studies reveals that in difficult tasks assessing for instance reasoning, task success is positively related to time on task, whereas in easy tasks, such as basic interactions with a computer interface, the relation is negative. Independent evidence for this line of reasoning comes from research suggesting that task difficulty within a given domain affects the association between time on task and task success. Neubauer (1990) investigated the correlation between the average time on task and the test score for figural reasoning tasks and found a zero correlation. However, for task clusters of low, medium, and high difficulty, he found negative, zero, and positive correlations, respectively. Similarly, in a recent study by Dodonova and Dodonov (2013), the strength of the negative correlation between time on task and accuracy in a letter sequence task tended to decrease with increasing task difficulty.

Time on Task Effects and Dual Processing Theory

An explanation for the heterogeneity of associations between time on task and task success may be provided by dual processing theory, which distinguishes between automatic and controlled mental processes (cf. Fitts & Posner, 1967; Schneider & Chein, 2003; Schneider & Shiffrin, 1977). Automatic processes are fast, proceduralized, and parallel; they require little effort and operate without active control or attention, whereas controlled processes are slow, are serial, require attentional control, and can be alternated quickly. Tasks are amenable to automatic processing due to learning only under consistent conditions, that is, rules for information processing including related information-processing components and their sequence are invariant (Ackerman, 1987). Learning under consistent conditions can be divided into three stages (cf. Ackerman & Cianciolo, 2000; Fitts & Posner, 1967). The first stage, when the individual acquires task knowledge and creates a production system (cf. Adaptive Control of Thought [ACT] theory; Anderson & Lebiere, 1998), is characterized by controlled processing. Automatic processing becomes more apparent in the second stage and dominates in the third stage. Thus, task performance is slow and error prone at the beginning of learning, but speed and accuracy increase as the strength of productions is increased through practice (Anderson, 1992).

Consequently, in domains and tasks that allow for automatic processing, a negative association between time on task and task success is expected. Well-practiced task completion is associated with both fast and correct responses. In contrast, a positive association is expected in domains and tasks that do not allow for a transition from controlled to automatic processing due to inconsistent processing rules and variable sequences of information processing. Taking more time to work carefully would positively impact task success. In line with this reasoning, Klein Entink et al. (2009)

showed that test effort in a reasoning test, that is, the extent to which a test taker cares about the result, is positively related to test-taking slowness (measured through time on task), which itself is positively related to skill (measured through task success).

Notably, dual processing theory suggests a dynamic interaction of automatic and controlled processing in that the acquisition of higher level cognition is enabled by and builds upon automatic subsystems (Shiffrin & Schneider, 1977). Basically, tasks within and between domains are assumed to differ with respect to the composition of demands that necessarily require controlled processing and those that can pass into automatic processing (Schneider, & Fisk, 1983). Similarly, for a particular task, individuals are assumed to differ in the extent to which the task-specific information-processing elements that can be automatized are actually automatized (e.g., Carlson, Sullivan, & Schneider, 1989). In the following two sections, we describe in detail how automatic and controlled processes may interact in the two domains considered, reading and problem solving.

Time on Task in Reading

Reading a text demands a number of cognitive component processes and related skills. Readers have to identify letters and words. Syntactic roles are then assigned to words, sentences are parsed for their syntax, and their meaning is extracted. Coherence must be established between sentences, and a representation of the propositional text base must be created, as well as a situation model of the text contents, integrated with prior knowledge (Kintsch, 1998). In addition, cognitive and metacognitive regulations might be employed. When text contents are learned, strategies of organization and elaboration will aid the learning process. These different cognitive component skills allow for a transition from controlled to automatic processing to different degrees. Processes such as phonological recoding, orthographic comparison, or the retrieval of word meanings from long-term memory are slow and error prone in younger readers but become faster and more accurate as reading skill acquisition progresses (Richter, Isberner, Naumann, & Neeb, 2013). Indeed, theories of reading such as the lexical quality hypothesis (Perfetti, 2007) claim that reading skill rests on reliable as well as quickly retrievable lexical representations. In line with this, text comprehension is predicted by the speed of access to phonological, orthographic, and meaning representations (e.g., Richter et al., 2012, 2013). Beyond the word level, the speed of semantic integration and local coherence processes are equally positively related to comprehension (e.g., Naumann, Richter, Christmann, & Groeben, 2008; Naumann, Richter, Flender, Christmann, & Groeben, 2007; Richter et al., 2012). As shown by longitudinal studies, accuracy in reading assessments during primary school approaches perfection, whereas reading fluency reflecting reading performance per time unit continues to increase across years of schooling (cf. Landerl & Wimmer, 2008). The high accuracy rates suggest that reading is already well automatized during primary school.

Following this line of reasoning, in reading tasks, a negative time on task effect might be expected. A number of reading tasks, however, require attentional cognitive processing to a substantial degree as well. For instance, readers might need to actively choose which parts of a text to attend to when pursuing a given reading goal (e.g., Gräsel, Fischer, & Mandl, 2000; Naumann et al., 2007, 2008;

Organisation for Economic Co-Operation and Development [OECD], 2011, chap. 3; Puntambekar & Stylianou, 2005). In the case of a difficult text, strategies such as rereading or engaging in self-explanations (e.g., Best, Rowe, Ozuru, & McNamara, 2005; McKeown, Beck, & Blake, 2009) are needed for comprehension. Also, in skilled readers, such processes require cognitive effort (Walczyk, 2000), and effort invested in strategic reading positively predicts comprehension (e.g., Richter, Naumann, Brunner, & Christmann, 2005; Sullivan, Gnedsdilow, & Puntambekar, 2011). This, however, will involve longer time spent on task.

Taken together, this means that in easy reading tasks, the potentially automatic nature of reading processes at the word, sentence, and local coherence level leads to a negative time on task effect (e.g., when reading a short and highly coherent linear text). As reading tasks become more difficult and readers need to engage in strategic and thus controlled cognitive processing, the negative time on task effect will be diminished or reversed.

Time on Task in Problem Solving

Problem solving is required in situations where a person cannot attain a goal by using routine actions or thinking due to barriers or novelty (e.g., Funke & Frensch, 2007; Mayer, 1992; Wirth & Klieme, 2003). Problem solving requires higher order thinking, the finding of new solutions, and sometimes interaction with a dynamic environment (Klieme, 2004; Mayer, 1994). In the present study, a specific concept of problem solving as defined for the PIAAC study is taken into account; it refers to solving information problems in technology-rich environments. That is, technology-based tools and information sources (e.g., search engines, Web pages) are used to solve a given problem by “storing, processing, representing, and communicating symbolic information” (OECD, 2009b, p. 8). Information problems in this sense (e.g., finding information on the Web fulfilling multiple criteria to take a decision) cannot be solved immediately and routinely. They require developing a plan consisting of a set of properly arranged subgoals and performing corresponding actions through which the goal state can be reached (e.g., identifying the need for information to be obtained from the Web, defining an appropriate Web search query, scanning the search engine results page, checking linked Web pages for multiple criteria, collecting and comparing information from selected Web pages, and making use of it in the decision to be taken). This differs, for instance, from solving logical or mathematical problems where complexity is determined by reasoning requirements but not primarily by the information that needs to be accessed and used (OECD, 2009b). Cognitive and metacognitive aspects of problem solving as assessed in PIAAC include setting up appropriate goals and plans to achieve the goal state. This includes monitoring the progress of goal attainment, accessing and evaluating multiple sources of information, and making use of this information (OECD, 2009b, p. 11).

Problem solving is a prototype of an activity that relies on controlled processing. Controlled processing enables an individual to deal with novel situations for which automatic procedures and productions have not yet been learned. Otherwise, the situation would not constitute a problem. Accordingly, Schneider and Fisk (1983) described skilled behavior in problem solving and strategy planning as a function of controlled processing. Notably, problem solving skill may also benefit from practice. The development of fluent component skills at the level of subgoals enables problem solvers to improve their strategies optimizing the problem solving process (see, e.g., Carlson, Khoo, Yaure, & Schneider, 1990).

General conceptualizations of (complex) problem solving conceive problem solving performance as consisting of knowledge acquisition including problem representation and the application of this knowledge to generate solutions (cf. Funke, 2001; Greiff, Wüstenberg, et al., 2013). Wirth and Leutner (2008) identified two simultaneous goals in the knowledge acquisition phase, that is, generating information through inductive search and integrating this information into a coherent model. Successful problem solvers move more quickly from identification to integration and thus will be able to invest time in advanced modeling and prediction (which provide the basis for successful knowledge application) rather than in low-level information processing.

Problem solving in technology-rich environments assumes two concepts, accessing information and making use of it, that seem similar to knowledge acquisition and application. However, there are differences in that, for instance, retrieving information (e.g., by means of a search engine) is not comparable to an inductive search for rules governing an unknown complex system. Nevertheless, the various notions of problem solving assume successive steps of controlled information processing that may benefit from fluent component skills.

Therefore, a positive effect of time on task on task success is expected for problem solving. Taking sufficient time allows for all serial steps to planned subgoals to be processed, as well as more sophisticated operations to be used and properly monitored regarding progress. Particularly for weak problems solvers, spending more time on a task may be helpful to compensate for a lack of automaticity in required subsystems (e.g., reading or computer handling processes).

Research Goal and Hypotheses

Our general research goal was to assess and investigate behavioral processes and their relation to task performance in computer-based assessment. More specifically, we determined the effect of time on task on the task result and the conditions that influence the strength and direction of this effect. For this, we used the computer-based assessment of reading and problem solving in the international large-scale study PIAAC, including log file data generated by the assessment system.

From a dual processing framework, we derived the general hypothesis that the relative degree of controlled versus automatic cognitive processing as required by a task, as well as the test taker's acquired skill level, determines the strength and direction of the time on task effect. The following three hypotheses address time on task effects across domains, task properties, and person characteristics. The fourth hypothesis aims at validating the interpretation of the time on task effect in problem solving by splitting up the global time on task into components that represent different steps of task solution and information processing.

Hypothesis 1: Time on task effect across domains. We expected a positive time on task effect for problem solving in technology-rich environment tasks. A negative time on task effect was expected for reading tasks because, in reading tasks, a number of component cognitive processes are apt for automatization. Problem solving, in contrast, by definition must rely on controlled processing to a substantial degree in each task.

Hypothesis 2: Time on task effect across tasks. Within domains, we expected the time on task effect to be moderated by task difficulty. Easy tasks can be assumed to be completed substantially by means of automatic processing, whereas difficult tasks evoking more errors require a higher level of controlled processing. Accordingly, we expected a positive time on task effect in problem solving to be accelerated with increasing task difficulty, and a negative time on task effect in reading to diminish with increasing task difficulty.
As our intepretation of the time on task effect focuses the way of cognitive processing, we additionally explored the potentially moderating role of the cognitive operation involved in each task as defined a priori by the PIAAC assessment framework (e.g., access in reading). More specifically, we investigated whether the task characteristic "cognitive operation" explains task difficulty and if so whether the time on task effect would depend on the presence of specific cognitive operations.

Hypothesis 3: Time on task effect across persons. For a given task, individuals are assumed to differ in the extent to which the information-processing elements that are amenable to automatic processing are actually automatized. Highly skilled individuals are expected to be in command of well-automatized procedures within task solution sub-systems that are apt to automatization (such as decoding in reading or using shortcuts to perform basic operations in a computer environment). We therefore expect the time on task effect to vary across persons. On the one hand, we predict that the time on task effect gets more positive for less skilled problem solvers and less negative for less skilled readers since they are expected to accomplish tasks with higher demands of controlled and strategic processing than skilled persons. For example, poor readers may rely on compensatory behaviors and strategies, especially when completing difficult tasks (see Walczyk, 2000). On the other hand, for skilled persons, we expect the inverse result, that is, due to a higher degree of routinized processing, the time on task effect gets less positive for skilled problem solvers and more negative for skilled readers.

Hypothesis 4: Decomposing time on task effect at task level. Computer-based assessment and especially the exploitation of log file data can help to further understand the task completion process. By moving from the global process measure of time on task to the underlying constituents, we can further validate the interpretation of the time on task effect. This is especially true for tasks requiring a complex sequence of stimulus interactions that can be reconstructed from a log file, giving insight into the accuracy and timing with which subgoals were being completed. In the present study, tasks assessing problem solving in technology-rich environments are highly interactive, requiring the operation of simulated computer and software environments or navigation in simulated Web environments. For a particular task, we expect that a positive time on task effect is confined to the completion of steps that are crucial for a correct solution (e.g., in a Web environment, visiting a page that presents information needed to give a correct response), whereas for others the effect is assumed to be negative (e.g., in a Web environment, visiting an irrelevant page). If this were the case, it would corroborate our assumption that it is the need for strategic and controlled allocation of cognitive resources that produces a positive time on task effect in problem solving or very difficult reading tasks.

Method

Sample

The PIAAC study initiated internationally by the OECD (cf. OECD, 2013; Schleicher, 2008) is a fully computer-based international comparative study assessing the competence levels of adults in 2011–2012. For the present study, data provided by GESIS–Leibniz Institute for the Social Sciences from the German PIAAC field test in 2010 were used. The target population consisted of all noninstitutionalized adults between the ages of 16 and 65 years (inclusive) who resided in Germany at the time of sample selection and were enrolled in the population register. For the field test in Germany, a three-stage sampling was used with probability sample of communities and individuals in five selected federal states. The within-household sample included in the present study comprised 1,020 individuals completing the computer-based PI-AAC assessment. Of these, 520 were male (50.98%) and 458 female (44.90%). For 42 participants, no gender information was available (4.12%). The average age was 39.40 years (SD = 13.30).

Instrumentation

Reading literacy. The PIAAC conceptual framework for reading literacy is based on conceptions of literacy from the International Adult Literacy Survey (IALS) conducted in the 1990s and the Adult Literacy and Life Skills Survey (ALL) conducted in 2003 and 2006 (see OECD, 2009a). It was extended for PIAAC to cover reading skill in the information age by including skills of reading in digital environments. More than half of the reading tasks were taken from the former paper-based adult literacy assessments IALS and ALL to link PIAAC results back to these studies. New tasks simulating digital (hypertext) environments were developed to cover the broadened construct including skills of reading digital texts. The tasks covered the cognitive operations “access and identify information,” “integrate and interpret information,” and “evaluate and reflect information” (see OECD, 2009a). The majority of tasks included print-based texts as used in previous studies (e.g., newspapers, magazines, books). Tasks representing the digital medium included, for instance, hypertext and environments such as message boards and chat rooms. Tasks are also varied with respect to the context (e.g., work/occupation, education and training) and whether they included continuous texts (e.g., magazine articles), noncontinuous texts (e.g., tables, graphs), or both.

In the PIAAC field test, 72 reading tasks were administered. For the present study, only those 49 tasks were used that entered the main study. To respond, participants were required to highlight text, to click a (graphical) element of the stimulus, to click a link, or to select a check box. As a sample, Figure 1 (upper panel) presents a screenshot from the first “Preschool Rules” task.

Respondents were asked to answer the question shown on the left side of the screen by highlighting text in the list of preschool rules on the right side. The question was to figure out the latest time that children should arrive at preschool. Thus, readers were required to access and identify information, the context was personal, and print text was presented.

Problem solving in technology-rich environments. This construct refers to using information and communication technology (ICT) to collect and evaluate information so as to communicate and perform practical tasks such as organizing a social activity, deciding between alternative offers, or judging the risks of medical treatments (OECD, 2009b). The framework (OECD, 2009b) defined multiple task characteristics that formed the basis for instrument development. The cognitive operations to be covered by the tasks were goal setting and progress monitoring, planning and self-organizing, acquiring and evaluating information, and making use of information. The technology dimensions included hardware devices (e.g., desktop or laptop computers), software applications (e.g., file management, Web browser, e-mail, spreadsheet), various commands and functions (e.g., buttons, links, sort, find), and multiple representations (e.g., text, numbers, graphics). Moreover, task development aimed at the variation of the task's purpose (e.g., personal, work/occupation), intrinsic complexity (e.g., the minimal number of actions required to solve the problem, the number of constraints to be satisfied), and the explicitness of the problem (implicit, explicit).

As defined by the framework (OECD, 2009b), tasks were developed in such a way that they varied in the number of required cognitive operations (e.g., acquiring and evaluating information), the number and kind of actions that have to be taken to solve the task in a computer environment, the inclusion of unexpected outcomes or impasses, and the extent to which the tasks were open-ended. A more difficult task simulating real-life problem solving would require several cognitive operations, multiple actions in different environments, unexpected outcomes, and the planning of multiple subgoals that may depend on each other. A corresponding sample task would be one in which the problem solver has to do a Web search on the Internet to access information, integrate and evaluate information from multiple online sources by using a spreadsheet, and then create a summary of the information to be presented at school by using a presentation software.

In the PIAAC field test, 24 problem solving tasks were administered. Of these tasks, only 13 were selected for the main study. For the present study, all available tasks were considered to obtain more reliable results on the correlation of effects varying across tasks. After excluding tasks with poor discrimination and tasks for which no score could be derived, 18 tasks were left. In the context of international large-scale assessments, further tasks may be dropped, especially if they show differential item functioning across participating countries. However, as we only used national data and did not aim at comparing countries, there was no need to consider task-by-country interactions. To give a response in the simulated computer environments, participants were required to click buttons, menu items, or links, to select from drop-down menus, to drag and drop, and so on.

As a sample, Figure 1 (lower panel) presents a screenshot from the task “Job Search.” Regarding cognitive operations, participants had to access and evaluate information and monitor criteria for

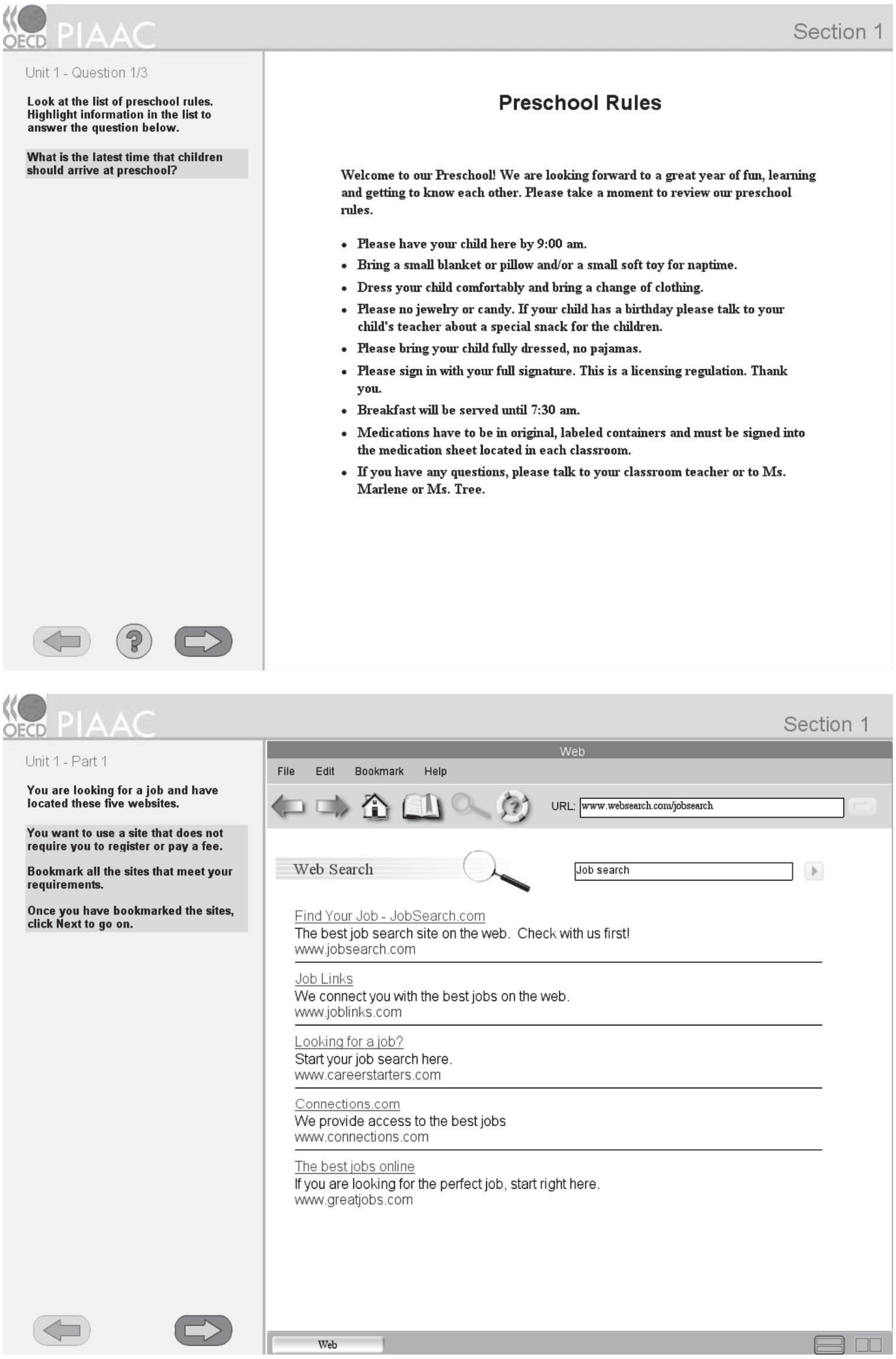

Figure 1. Sample tasks: reading literacy task “Preschool Rules” (upper panel); problem solving in technology-rich environments task “Job Search” with only the start page showing the search engine results depicted, not the linked pages (lower panel). OECD = Organisation for Economic Co-Operation and Development; PIAAC = Programme for the International Assessment of Adult Competencies.

constraint satisfaction within a simulated job search. Thus, the task's purpose was occupational. Starting from a search engine results page, the task was to find all the sites that do not require users to register or pay a fee and to bookmark these sites. Regarding the explicitness of the problem, instructions did not directly tell participants the number of sites they must locate, but evaluation criteria were clearly stated. To solve the task, single actions of evaluation had to be repeated for each website; for a target page, multiple constraints needed to be satisfied. Both characteristics determined intrinsic complexity. As regards software applications and related commands, the task was situated in a simulated Web environment that included tools and functionality similar to those found in real-life browser applications, that is, clickable links, back and forward buttons of the browser, and a bookmark manager that allowed one to create, view, and change bookmarks. The opening page presented the task description on the left side and the results of the Web search engine, that is, clickable links and brief information about the linked page, on the right side of the screen. From this search engine results page, participants had to access the hypertext documents connected via hyperlinks to locate and bookmark those websites that meet the search criteria.

Design and Procedure

A rotation design was used to form 21 booklets resulting in an effective sample size for reading literacy of 113 to 146 responses per task and for problem solving in technology-rich environments of 140 to 191 responses per task.

Data were collected in computer-assisted personal interviews. Interviewers went to the participants' households to conduct the interview in person. First, participants completed a background questionnaire, and then the interviewer handed the notebook to the participant for completion of the cognitive tasks. There was no global time limit, that is, participants could take as long as they needed. Participants only completed the computer-based tasks if they were sufficiently ICT literate, which was tested by ICT tasks requiring basic operations such as highlighting text by clicking and dragging. In case of nonsufficient ICT literacy, a paper-based assessment was administered. In the computer-based part, participants were randomly assigned to booklets including reading literacy, numeracy, and problem solving tasks. For the present study, only data from the computer-based assessment of reading literacy and problem solving were included.

Statistical Analyses

Modeling approach. The generalized linear mixed model (GLMM) framework (e.g., Baayen, Davidson, & Bates, 2008; De Boeck et al., 2011; Doran, Bates, Bliese, & Dowling, 2007) was used to investigate the role of time on task in reading and problem solving (Hypotheses 1–3). A linear model consists of a component η_{pi}, representing a linear combination of predictors determining the probability of person *p* for solving task *i* correctly. The predictors' weights are called effects. Modeling mixed effects means to include both random effects and fixed effects. Fixed effects are constants across units or groups of a population (e.g., tasks, persons, classrooms), whereas random effects may vary across units or groups of a population (cf. Gelman, 2005). The generalized version of the linear mixed model accommodates also categorical response variables. In measurement models of item response theory, for instance, the effect of each item or task i on the probability of obtaining a correct response is typically estimated as a fixed effect representing the task's difficulty or easiness. The effect of person p is usually modeled as random, that is, as an effect which may vary across persons and for which the variance is estimated. The variance of this random effect represents the variability of skill across persons.

The GLMM incorporating both random effects, **b**, and fixed effects, **β**, can be formulated as follows: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ (e.g., Doran et al., 2007). In this model, **X** is a model matrix for predictors with fixed weights included in vector **β**, and **Z** is a model matrix for predictors with random weights included in vector **b**. The distribution of the random effects is modeled as a multivariate normal distribution, $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, with **Σ** as the covariance matrix of the random effects. The continuous linear component η_{pi} is linked to the observed ordered categorical response Y_{pi} (correct vs. incorrect) by transforming the expected value of the observed response, that is, the probability to obtain a correct response π_{pi}. When using the log-transformed odds ratio (log-odds), the logit link function follows:

$\eta_{pi} = \ln(\pi_{pi}/(1- \pi_{pi}))$ (cf. De Boeck et al., 2011).

In the present study, to address the research question of whether the strength of the time on task effect is correlated with the easiness of tasks, the effects of both persons and tasks were defined as random intercepts (cf. random person random item model; De Boeck, 2008). A fixed intercept, β_0, is estimated additionally, which is the same for all participants and tasks.

A baseline Model M0 was obtained by specifying an item response model (1PL or Rasch model) with task and person as random intercepts and by adding the time on task as person-by-item predictor with a fixed effect β_1. Model M0 serves as parsimonious reference model that is compared with more complex models including further fixed and/or random effects: η_{pi} = (intercept β_0) + (individual skill b_{0p}) + (relative easiness b_{0i}) + β_1 (time on task t_{pi}).

In the following analyses, this model is systematically extended by adding further predictors. For example, the predictor (time on task t_{pi}) with the random weight b_{1i} is added, providing the variance of the by-task adjustment b_{1i} to the fixed time on task effect β_1. As the by-task adjustment, b_{1i}, and task easiness, b_{0i}, are tied to the same observational unit, that is, task i, their association is also estimated. This correlation can be used to test whether the strength of the time on task effect linearly depends on task difficulty (as claimed by Hypothesis 2). Figure 2 shows the path diagram of Model M1, which is Model M0 extended by the predictor (time on task t_{pi}) with a random weight across tasks, b_{1i} (cf. the graphical representations of GLMMs by De Boeck & Wilson, 2004). In Model M1, there is a fixed time on task effect, β_1, representing the average time on task effect. However, it is adjusted by task by adding the weight b_{1i}, which allows the time on task effect to vary across tasks as indicated by subscript i. The other models under consideration can be derived in a similar fashion by adding random effects adjusting the time on task effect by cognitive operation (Model M2, cf. Hypothesis 2), by person (Model M3, cf. Hypothesis 3), or by task and person (Model M4, integrating Hypothesis 2 and Hypothesis 3).

To clarify whether the introduction of further random components into the model significantly improves model fit, model comparison tests were conducted. For comparing nested models, the likelihood ratio (LR) test was used, which is appropriate for inference on random effects (Bolker et al., 2009). The test statistic, that is, twice the difference in the log-likelihoods, is approximately χ^2 distributed with degrees of freedom equal to the number of extra parameters in the more complex model. The LR test is problematic when the null hypothesis implies the variance of a random effect to be zero; this means that the parameter value is on the boundary of the parameter space (boundary effect; cf. Baayen et al., 2008; Bolker et al., 2009; De Boeck et al., 2011). Using the chi-square reference distribution increases the risk of Type II errors; therefore, the LR test has to be considered as a conservative test for variance parameters.

For the analysis at the task level (Hypothesis 4), logistic regression was used to predict task success by the time taken on individual steps of the task completion sequence.

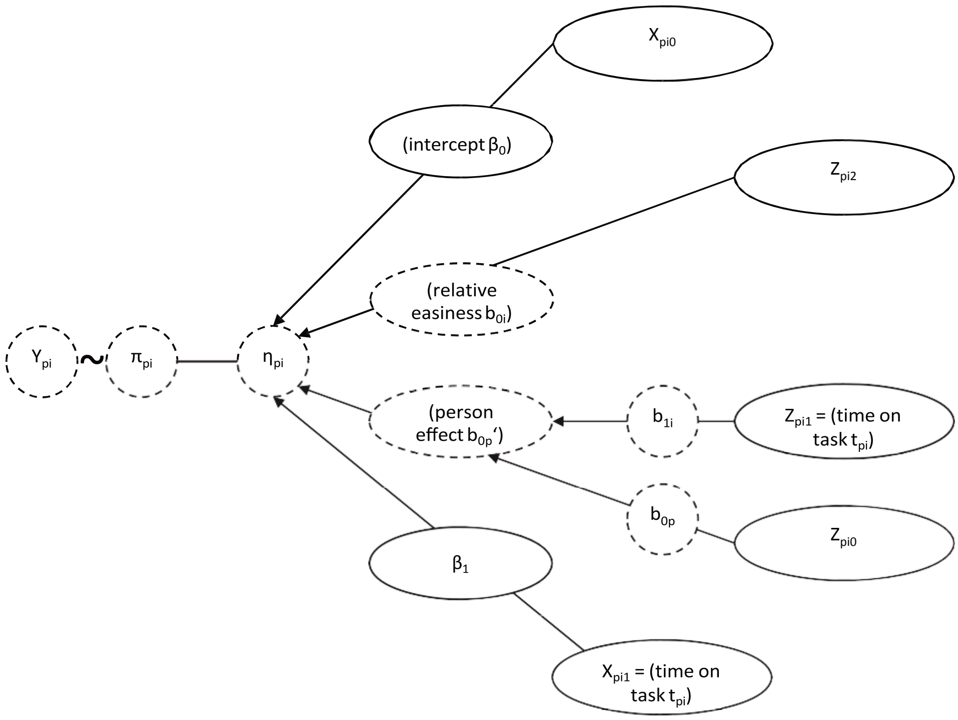

Figure 2. Graphical representation of model M1 showing how the probability to obtain a correct response, η_{pi}, is affected by a general intercept, β_0, the relative task easiness, b_{0i}, and individual skill, b_{0p}. Moreover, there is a time on task effect consisting of a fixed part, β_1, as well as random part, b_{1i}, which means that the time on task effect may vary across tasks *i*.

Interpreting the effect of time on task in the GLMM. The "fundamental equation of RT modeling" (van der Linden, 2009, p. 259) assumes that the response time (RT; time on task) of person *p* when completing task *i* depends both on the person's speed τ_p and the task's time intensity λ_i. Accordingly, the expected value of the (log-transformed) response time can be defined as follows: $E(\ln(t_{pi})) = \lambda_i - \tau_p$ (cf. van der Linden, 2009). This implies that the effect of time on task reflects both the effect of the person and the task component.

When the effect of time on task is introduced as an overall fixed effect β_1, as in Model M0, this effect would reflect the association between time on task and the log-odds ratio of the expected response. This association could not be interpreted in a straight-forward way, as it depends not only on the

correlation between underlying person-level parameters, that is, skill and speed, but also on the correlation of corresponding item parameters, that is, difficulty and time intensity (see van der Linden, 2009).[2] However, when modeling the effect of time on task as an effect random across tasks (Hypothesis 1), groups of tasks supposed to be homogeneous (Hypothesis 2), or individuals (Hypothesis 3), the influences from the task and person levels can be disentangled.
A time on task effect random across tasks is obtained by introducing the by-task adjustment b_{1i} to the fixed time on task effect β_1. The time on task effect by task results as $\beta_1 + b_{1i}$. Thereby, time on task is turned into a person-level covariate varying between tasks. That is, given a particular task with certain time intensity, variation in time on task is only due to differences in persons' speed (plus residual). This allows us to interpret time on task as an task-specific speed parameter predicting task success above and beyond individual skill.
A by-person random time on task effect means to adjust the fixed time on task effect β_1 by the person-specific parameter b_{1p}, resulting in the time on task effect $\beta_1 + b_{1p}$. The fixed effect shows a constant as subscript, whereas the random effect is provided additionally with p as subscript indicating that the effect may vary across persons p. Given a particular person working at a certain speed level, variation in time on task is only due to differences in the tasks' time intensity (plus residual). This means that time on task can be conceived of as a task-level covariate that is specific to persons and predicts task success above and beyond task easiness.

Trimming of time data. As a preparatory step for data analysis, the (between-person) time on task distribution of each task was inspected for outliers. The middle part of a time on task distribution was assumed to include the observations that are most likely to come from the cognitive processes of interest. To exclude extreme outliers in time on task and to minimize their effect on analyses, observations two standard deviations above (below) the mean were replaced by the value at two standard deviations above (below) the mean. As even a single extreme outlier can considerably affect mean and standard deviation, time on task values were initially log-transformed, which means that extremely long time on task values were pulled to the middle of the distribution. With this trimming approach, 4.79% of the data points in reading literacy and 4.67% in problem solving were replaced. Transforming a covariate may have an impact on estimated parameters of the linear mixed model (for linear transformations, see, e.g., Morrell, Pearson, & Brant, 1997). Therefore, we conducted the analyses also without log-transforming the time on task variable. As we obtained the same result pattern, we report the analyses with log transformation only. Results obtained with the untransformed data are available from the first author upon request.

Statistical software. For estimating the presented GLMMs, the lmer function of the R package lme4 (Bates, Maechler, & Bolker, 2012) was applied. The R environment (R Core Team, 2012) was also used to conduct logistic regression analyses.

[2] We thank an anonymous reviewer who advised us to consider this issue.

Results

Difficulty of Tasks

To compare the difficulty of problem solving tasks and reading literacy tasks, the baseline Model M0 was tested for both domains without the time on task effect. For reading literacy, an intercept of β_0 = 0.61 (z = 3.21, p < .01) was obtained; it represents the marginal log-odds for a correct response in a task of average easiness completed by a person of average skill; the corresponding probability was 64.68%. For problem solving, the result was β_0 = -0.72 (z = -2.37, p < .01), indicating that the probability of a correct response was on average only 32.68%, that is, problem solving tasks were much harder than reading literacy tasks. Figure 3 shows the densities of the estimated task easiness parameters for reading literacy tasks (upper panel) and problem solving tasks (lower panel). Task easiness values were obtained by adding the intercept β_0 and the random task intercept (relative easiness b_{0i}). The proportion of correct responses, *p*, ranged for reading literacy from 12.41% to 96.92% and for problem solving from 11.86% to 77.49%.

Time on Task Effect by Domain (Hypothesis 1)

For testing Hypotheses 1 and 2, Model M0 was extended to Model M1 by adding the by-task random time on task effect b_{1i}: η_{pi} = (intercept β_0) + (individual skill b_{0p}) + (relative easiness b_{0i}) + β_1 (time on task t_{pi}) + b_{1i} (time on task t_{pi}).

To address Hypothesis 1 regarding the time on task effect by domain, the fixed time on task effects β_1, as specified in Model M1 (see also Figure 2), were compared between reading literacy and problem solving.

Reading literacy. Table 1 provides an overview of the results. For reading literacy, a negative and significant time on task effect of β_1 = -0.61 (z = -4.90, p < .001) was found. Thus, for a reading literacy task of average difficulty, correct responses were associated with shorter times on task, whereas incorrect responses were associated with longer times on task.

Problem solving. For problem solving, a positive and significant time on task effect of β_1 = 0.56 (z = 2.30, p = .02) was estimated. Thus, for a problem solving task of average difficulty, correct responses were associated with longer times on task and vice versa. These findings give support to Hypothesis 1.

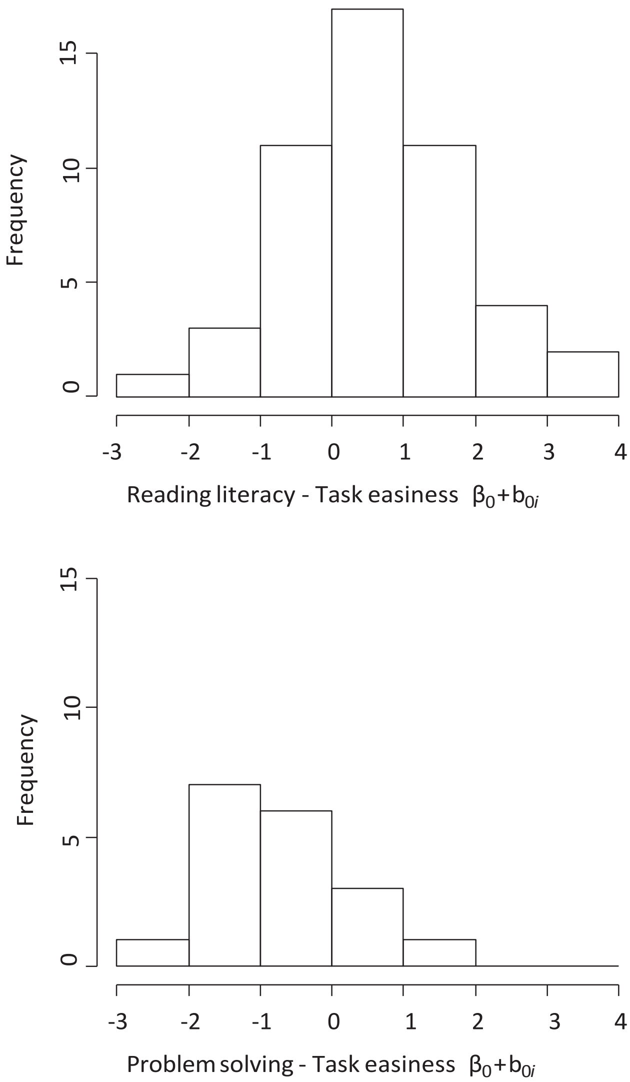

Figure 3. Distribution of estimated task easiness parameters for reading literacy (upper panel) and problem solving in technology-rich environments (lower panel). On average, reading literacy tasks were easier than problem solving tasks.

Time on Task Effect by Task (Hypothesis 2)

If the assumption holds that task difficulty moderates the time on task effect, a relation between task easiness and the strength of the time on task effect should be observable within a domain. To test Hypothesis 2, the variances of the by-task adjustments to the fixed time on task effects and their correlations with task easiness, as estimated through Model M1, were inspected for both domains under consideration.

Reading literacy. For reading literacy, the variability of the by-task adjustment was estimated to be $\mathrm{Var}(b_{1i}) = 0.55$. This means that for reading literacy, the time on task effect varied across tasks. Most importantly, the by-task time on task effect and intercept were negatively correlated, $\mathrm{Cor}(b_{0i}, b_{1i}) = -.39$. That is, the overall negative time on task effect became even stronger in easy tasks but was attenuated in difficult tasks. The upper left panel in Figure 4 illustrates how the time on task effect in reading literacy was adjusted by task. To test whether the model extension improved the model's goodness of fit, we compared the nested Models M0 and M1. The difference test showed that Model M1 fitted the data significantly better than Model M0, $\chi^2(2) = 77.65$, $p < .001$. To test whether the correlation parameter was actually needed to improve model fit, that is, to test the significance of the correlation, Model M1 was compared to a restricted version (Model M1r), which did not assume a correlation between by-task time on task effect and by-task intercept. The model difference test suggested that the unrestricted version of Model M1 had a better fit to the data than the restricted version, $\chi^2(1) = 5.16$, $p = .02$. Thus, the negative correlation between the by-task adjustment of the time on task effect and the random task intercept (i.e., task easiness) was also significant.

Problem solving. For problem solving, the variance of the by-task adjustment to the fixed effect of time on task was estimated as $Var(b_{1i}) = 0.89$. Thus, for problem solving in technology-rich environments, the time on task effect varied across tasks. The correlation between the by-task adjustment to the time on task effect and task easiness was negative as for reading literacy, $Cor(b_{0i}, b_{1i}) = -.61$. That is, the overall positive time on task effect became even stronger in hard-to-solve tasks but was attenuated in easy-to-solve tasks. Figure 4 (upper right panel) illustrates how the time on task effect in problem solving was adjusted by task. The model difference test, comparing the nested Models M0 and M1, clearly showed that adding the random time on task effect in Model M1 improved the model fit, χ^2 (2) = 73.99, $p < .001$. Moreover, comparing Model M1 with a restricted version (Model M1r) without a correlation between the by-task time on task effect and the random task intercept revealed that the correlation was significant, $\chi^2(1) = 6.50$, $p = .01$.

All together, these results give clear support to Hypothesis 2. In a domain where task solution cannot rely on automatic processes such as problem solving, the already positive time on task effect was substantially increased in tasks that were especially difficult. In a domain where rapid automatic processing can account for a substantial part of the task solution process such as reading, an already negative time on task effect became even stronger in easier tasks but diminished in more difficult tasks.

Time on Task Effect by Cognitive Operation

An alternative explanation for the variability of the time on task effect between tasks refers to differences in the required cognitive operations. That is, tasks being homogeneous with respect to cognitive operations would show similar time on task effects. To test whether the presence of different cognitive operations as detailed by the respective frameworks affects the time on task effect, we extended Model M0 to the following Model M2 by introducing the cognitive operation c required in a task as a categorical task-level predictor and as a factor moderating the time on task effect, which is represented by the random weight b_{1c}: η_{pi} = (intercept β_0) + (individual skill b_{0p}) + (relative easiness b_{0i}) + β_1 (time on task t_{pi}) + (cognitive operation b_{0c}) + b_{1c} (time on task t_{pi}).

Table 1
Overview of Main Model Parameters on the Time on Task Effect and Model Comparison Tests

Domain	Research question/hypothesis	Model	Time on task effect random across	χ^2 of model difference test (*df* in parentheses)	Fixed-effect β1	Variance of random effect	Correlation of random effects
Reading literacy	Baseline model	M0			-0.55***	—	—
	Testing Hypotheses 1 and 2: Time on task effect by domain and task	M1	Tasks		-0.61***	0.55	-.39
	Comparison with baseline model	M1 vs. M0		77.65 (2)***			
	Restricted model without random effect correlation	M1r	Tasks		-0.59***	0.54	—
	Comparison with unrestricted model	M1 vs. M1r		5.16 (1)*			
	Exploring the time on task effect by cognitive operation	M2	Cognitive operations		-0.51***	0.003	-1.00
	Restricted model without random time on task effect across cognitive operations	M2r			-0.55***	—	—
	Comparison with unrestricted model	M2 vs. M2r		0.79 (1), ns			
	Testing Hypothesis 3: Time on task effect by person	M3	Persons		-0.65***	0.14	-.65
	Comparison with baseline model	M3 vs. M0		15.09 (2)**			
	Restricted model without random effect correlation	M3r	Persons		-0.57***	0.09	—
	Comparison with unrestricted model	M3 vs. M3r		12.85 (1)**			
	Integrated model: Time on task effect by task and person	M4	Tasks Persons		-0.69**	0.64 0.23	-.52 -.78
	Comparison with baseline model	M4 vs. M0		106.14 (4)***			
Problem solving	Baseline model	M0			0.49***	—	—
	Testing Hypotheses 1 and 2: Time on task effect by domain and task	M1	Tasks		0.56*	0.89	-.61
	Comparison with baseline model	M1 vs. M0		73.99 (2)***			
	Restricted model without random effect correlation	M1r	Tasks		0.54*	0.87	—
	Comparison with unrestricted model	M1 vs. M1r		6.50 (1)*			
	Testing Hypothesis 3: Time on task effect by person	M3	Persons		0.51***	0.22	-.79
	Comparison with baseline model	M3 vs. M0		5.98 (2)†			
	Restricted model without random effect correlation	M3r	Persons		0.49***	0.12	—
	Comparison with unrestricted model	M3 vs. M3r		5.98 (1)*			
	Integrated model: Time on task effect by task and person	M4	Tasks Persons		0.56*	0.89 0.11	-.63 -.76
	Comparison with baseline model	M4 vs. M0		76.77 (4)***			

Note. Dashes indicate that a parameter was not included in the model. M = Model; r = restricted. ns = not significant.
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Reading literacy. For reading literacy, the PIAAC framework assumes three broad aspects of cognitive operation, access and identify information, integrate and interpret information, and evaluate and reflect information. In a first step, we tested an explanatory item response model with random person and task effects as well as the effect of cognitive operation. For the three aspects of cognitive operations, the intercepts of 1.07 (z = 4.72, p < .01), 0.00 (z = 0.00, p = 1.00), and 0.08 (z = 0.19, p = .85) were estimated. The probabilities of a correct response corresponding to these intercepts were 74.50%, 50.01%, and 51.96%. Access tasks were thus relatively easy, whereas integrate and evaluate tasks show quite the same level of medium difficulty; by introducing cognitive operation as an explanatory variable of task easiness, the variance of task easiness, Var(b_{0i}), decreased from 1.52 to 1.24, which corresponds to R^2= .20.
To investigate whether the influence of time on task on task success varies across cognitive operations, Model M2 was tested. The obtained variance of the by-cognitive operation adjustment to the time on task effect was only Var(b_{1c}) = 0.003. Moreover, the correlation with the corresponding intercept was Cor(b_{0c}, b_{1c}) = -1.00, indicating overparameterization of the model. Model M2 was compared with a restricted model including no time effect varying across cognitive operations (Model M2r); there was no significant improvement of model fit, $\chi^2(2)$ = 0.79, p = .67. Thus, the time on task effect did not vary across cognitive operations.

Problem solving. The time on task effect was not further investigated with respect to cognitive operations for two reasons. First, there was only a small set of 18 tasks available. Second, each of the problem solving tasks explicitly included multiple cognitive operations from a set of four dimensions, that is, goal setting and progress monitoring, planning and self-organizing, acquiring and evaluating information, and making use of information, as defined by the PIAAC assessment framework (OECD, 2009b, p. 10). Given the constraints of a large-scale assessment, PIAAC only aimed at an overall indicator of problem solving. Our analyses would require a more fine-grained measure with a broad set of indicators for the various underlying cognitive operations. Although, for each task, one operation is assumed to be dominant, other operations might also be involved. For instance, the PIAAC framework maps the sample task "Job Search" to the cognitive operations of access and evaluating information as well as monitoring criteria for constraint satisfaction. There were only two more tasks that showed a comparable set of assumed cognitive operations, whereas in other tasks the requirement of accessing information was combined with a different additional demand, for example, communicating information. Thus, it was not possible to form subgroups with a sufficient number of tasks being homogeneous in the assumed composition of required cognitive operations.

Time on Task Effect by Person (Hypothesis 3)

On the person level, we assumed that the effect of time on task varies across the individual skill level. To test Hypothesis 3, we extended Model M0 to Model M3 by adding a random time on task effect, b_{1p}, representing the variation across individuals: η_{pi} = (intercept β_0) + (individual skill b_{0p}) + (relative easiness b_{0i}) + β_1 (time on task t_{pi}) + b_{1p} (time on task t_{pi}).

Reading literacy. For reading literacy, the variance of the by-person adjustment was Var(b_{1p}) = 0.14. Thus, for reading literacy, the time on task effect varied across persons. Most importantly, a correlation between the by-person time on task effect and by-person intercept of Cor(b_{0p}, b_{1p}) = -.65 was estimated. That is, the overall negative time on task effect became stronger in able readers but was attenuated in poor readers. The bottom left panel in Figure 4 illustrates how the time on task effect adjusted by person linearly decreases in more able persons. To clarify whether the liberal Model M3 better fitted the data, we compared the nested Models M0 and M3. The model difference test revealed that Model M3 fitted the data significantly better than Model M0, $\chi^2(2) = 15.09$, $p < .01$.

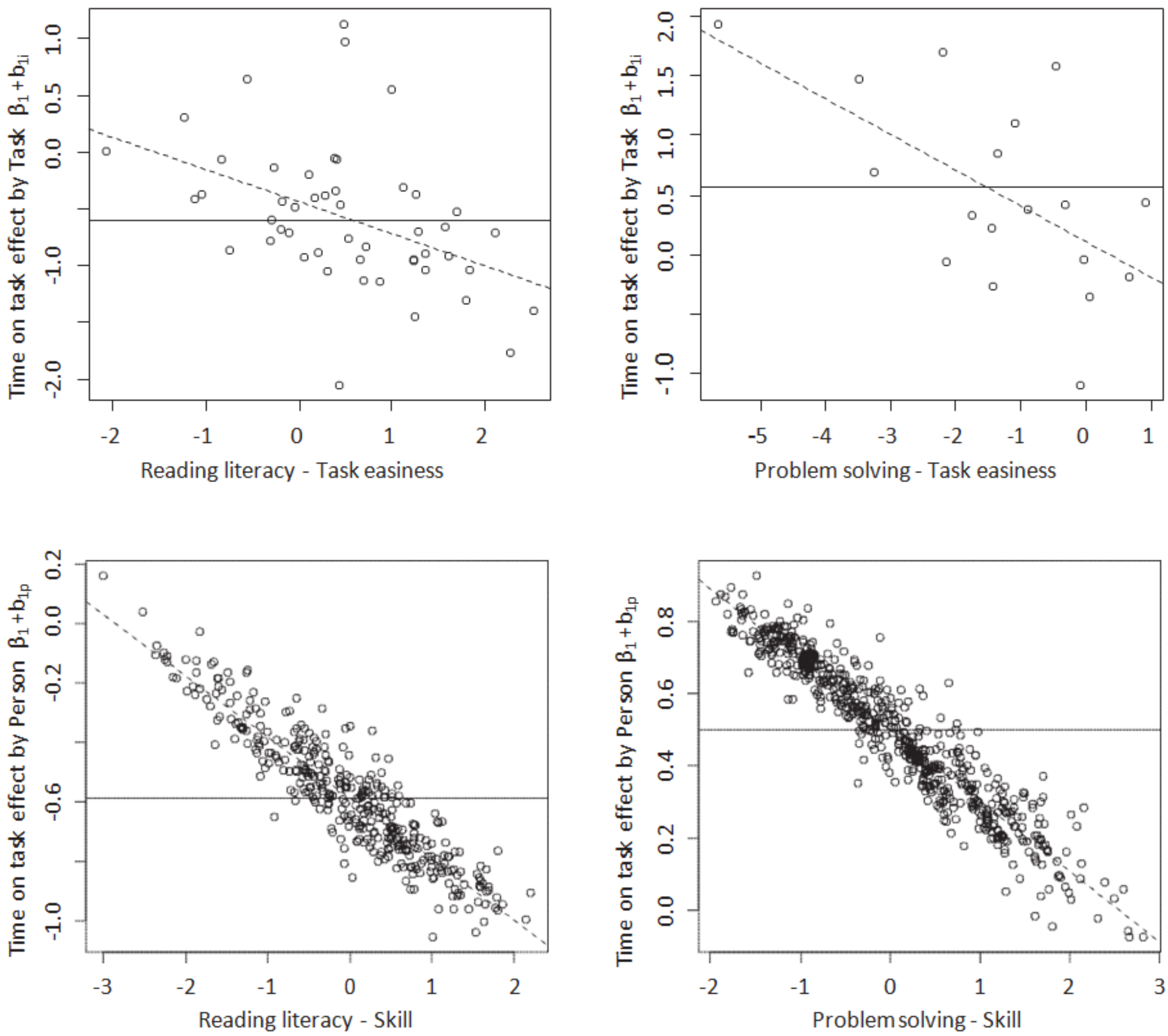

Figure 4. Upper row: Time on task effect by task for reading literacy (left panel) and problem solving in technology-rich environments (right panel). The solid line indicates the fixed time on task effect; the dots show how it is adjusted by task. For difficult tasks, the time on task effect gets more positive, whereas it gets more negative for easy tasks. Lower row: Time on task effect by person for reading literacy (left panel) and problem solving in technology-rich environments (right panel). The solid line indicates the fixed time on task effect; the dots show how it is adjusted by person. For less able individuals, the time on task effect gets more positive, whereas for able persons, it gets more negative.

To test whether the correlation parameter is required to improve model fit, that is, to test the significance of the correlation, Model M3 was compared with a restricted version (Model M3r) without the correlation between by-person time on task effect and intercept. The model difference test revealed that Model M3 without restrictions was the better fitting model, $\chi^2(1) = 12.85$, $p < .01$.

Problem solving. Similar results were obtained for problem solving. The variance of the by-person adjustment to the fixed effect of time on task was $\mathrm{Var}(b_{1p}) = 0.22$. Thus, for problem solving in technology-rich environments, the time on task effect varied across persons. The correlation between the by-person adjustment of the time on task effect and the by-person intercept (individual skill) was again negative and substantial, $\mathrm{Cor}(b_{0p}, b_{1p}) = -.79$. That is, the overall positive time on task effect became even stronger in poor problem solvers but was attenuated in able problem solvers (see the bottom right panel in Figure 4). The difference test comparing Model M3 including the random time on task effect with the baseline Model M0 was almost significant, $\chi^2(2) = 5.98$, $p = .05$. Finally, comparing Model M3 with a restricted version (Model M3r) without a correlation between by-task time on task effect and intercept revealed that the correlation was significant, $\chi^2(1) = 5.98$, p=.01.

Integrated Model: Time on Task Effect by Task and Person

As assumed in Hypotheses 2 and 3, the previous results indicate that task difficulty and individual skill level have an influence on the strength and direction of the time on task effect. The final Model M4 integrates both the by-task and the by-person adjustments to the time on task effect. The results found for Models M1 and M3 were perfectly reproduced in the following Model M4: η_{pi} = (intercept β_0) + (individual skill b_{0p}) + (relative easiness b_{0i}) + β_1 (time on task t_{pi}) + b_{1i} (time on task t_{pi}) + b_{1p} (time on task t_{pi}).

Reading literacy. For reading literacy, the time on task effect was estimated to be $\beta_1 = -0.69$ ($z = -5.16$, $p < .01$). The variance of the by-task adjustment to the time on task effect was $\mathrm{Var}(b_{1i}) = 0.64$, and that of the by-person adjustment was $\mathrm{Var}(b_{1p}) = 0.23$, that is, the time on task effect varied across both reading tasks and readers. Moreover, the time on task effect varied systematically in that the adjustments were linearly related to task easiness and individual skill level, respectively, as expected. The correlation between easiness of reading tasks and by-task adjustment was $\mathrm{Cor}(b_{0i}, b_{1i}) = -.52$, and the correlation between individual skill and by-person adjustment was $\mathrm{Cor}(b_{0p}, b_{1p}) = -.78$. The difference test showed that model M4 fit the data significantly better than model M0, $\chi^2(4) = 106.14$, $p < .001$.

The curves in Figure 5 (upper panel) indicate how for a given reader and reading task the probability for a correct response depends on time on task. The range of the time on task axis represents the empirical range of time on task in the selected tasks. The slope of the curves resulted from adding up the time on task effect and the adjustments to the time on task effect by task and by person. When considering a proficient reader (skill level of $b_{0p} = 1.61$) and an easy reading task (easiness of $b_{0i} = 1.89$), that is, a reading situation of low demand, the unadjusted negative effect of -.69 became much stronger, resulting in a negative time on task effect of -1.90 (plus line). However, in a situation of high

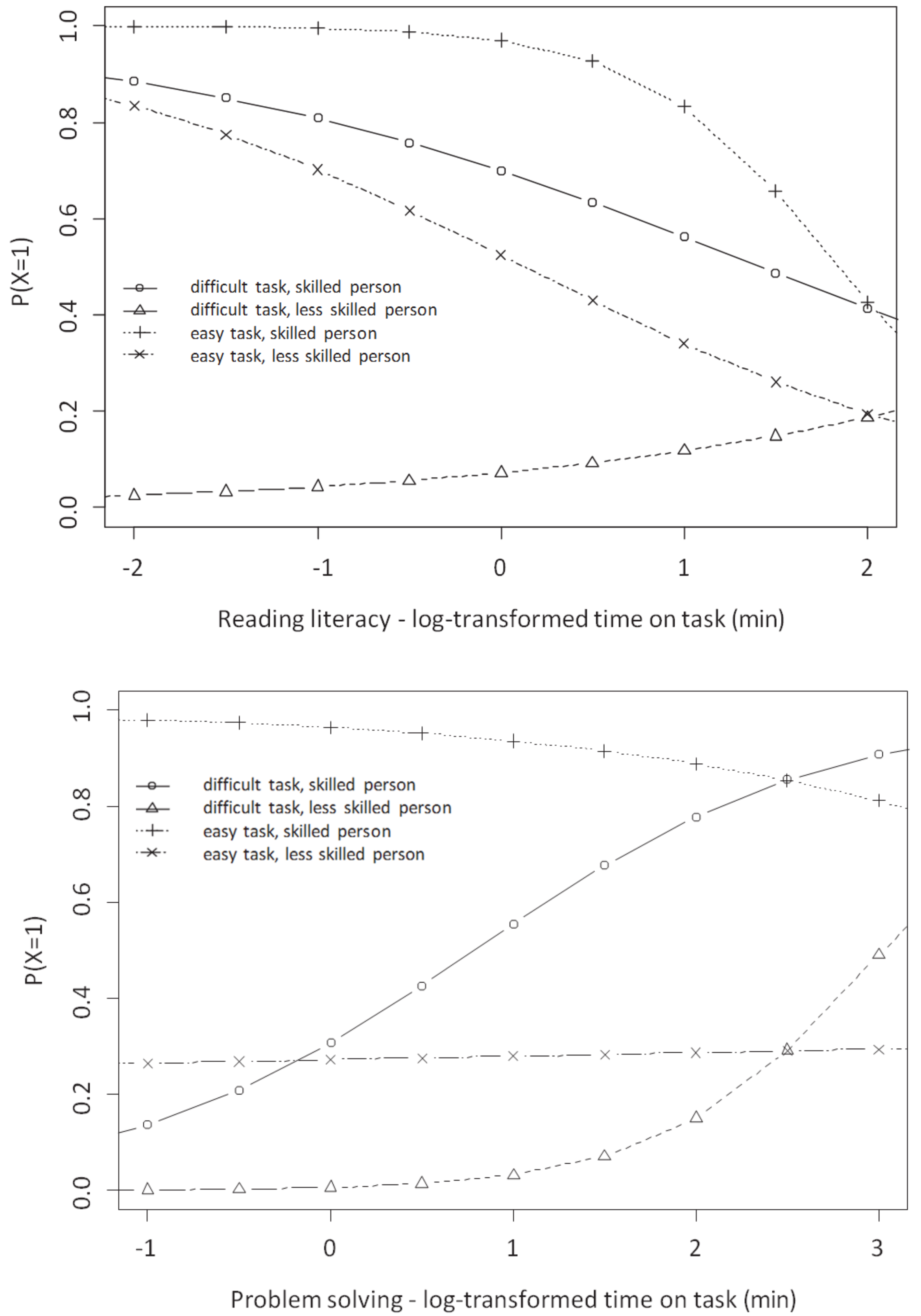

Figure 5. Time on task effect by task and skill level for reading. literacy (upper panel) and problem solving in technology-rich environments (lower panel). For combinations of two tasks (easy vs. hard) with two persons (less able vs. able), the probability of obtaining a correct response is plotted as a function of time on task.

demand, where a difficult reading task (easiness of b_{0i} = -0.77) was completed by a poor reader (skill level of b_{0p} = -1.79), the curve's slope was no longer negative but even slightly positive, that is, 0.55 (triangle line). In situations of medium demand, that is, a poor reader completing an easy task or an able reader completing a difficult task, the curves' slopes are in-between.

Problem solving. In the integrated model, a positive time on task effect of β_1 = 0.56 (z = 2.26, p = .02) was obtained. The variance of the by-task adjustment to the time on task effect was Var(b1i) = 0.89, and that of the by-person adjustment was Var(b_{1p}) = 0.11. The correlation between easiness of problem solving tasks and the by-task adjustment to the time on task effect was Cor(b_{0i}, b_{1i}) = -.63, and the correlation between individual skill level and the by-person adjustment to the time on task effect was Cor(b_{0p}, b_{1p}) = -.76. Again, the model comparison test indicated that model M4 fit the data significantly better than model M0, $\chi^2(4) = 76.77$, $p < .001$.
The bottom panel in Figure 5 shows the probability of obtaining a correct response as a function of the time on task for two selected tasks completed by two selected persons. In a situation of low demand, that is, a proficient problem solver (skill level of b_{0p} = 2.63) completing an easy task (easiness of b_{0i} = -0.67), the time on task effect decreases dramatically and becomes even negative and was estimated as -0.62 (+ line in Figure 5). However, in the situation of high demand where a difficult task (easiness of b_{0i} = -3.44) is completed by a poor problem solver (skill level of b_{0p} = -1.66), the positive time on task effect of .56 becomes much stronger and was estimated as 1.69 (Δ line in Figure 5). If the demand is medium, that is, a less able person completes an easy task or an able person completes a difficult task, the curves' slopes are in-between.

Taken together, these results indicate that positive time on task effects are observed especially in highly demanding situations, where not-so-skilled readers or problem solvers are confronted with a difficult task. Presumably, they can partly compensate for task demands by allocating cognitive resources. If this interpretation holds true, differential time on task effects should be observable on a within-task level as well. Specifically, if it is the strategic allocation of processing time that drives a positive time on task effect in problem solving tasks and difficult reading tasks being encountered by poor readers, on a within-task level the positive time on task effect should be confined to the processing of task-relevant parts of the stimulus. We tested this hypothesis as a last step.

Decomposing the Time on Task Effect at the Task Level (Hypothesis 4)

Using fine-grained time information extracted from log files, we decomposed the global time on task into several components that reflect particular steps of task solution. This was done at the task level for the problem solving task "Job Search," which required screening a search engine results page (see Figure 1, lower panel) and visiting multiple linked Web pages. Two of five Web pages in this task meet the criteria specified in the instruction and have to be bookmarked to obtain a correct response. In Hypothesis 4, spending more time on the two target pages was expected to indicate strategic behavior associated with a higher probability of successful task completion. In contrast, a negative effect was assumed for spending time on the search engine results page, which did not provide any hints about the target pages. For the time spent on nontarget pages, a negative effect was also expected.

First, logistic regression was used to predict the task success by time on task. The sample size for this analysis was 182. This analysis revealed a nonsignificant time on task effect of -0.29 (z = -0.59, p = .55). As a second step, task success was predicted by the time spent on the search engine results page, the time spent on the two relevant Web pages, and the time spent on the three irrelevant Web pages. The obtained effect for time spent on the relevant Web pages was positive and significant as expected, 0.96 (z = 2.53, p = .01), that is, spending more time on the target pages for evaluating the accessed information and monitoring the multiple criteria for constraint satisfaction was associated with a higher probability of achieving a correct response. In contrast, for the time spent on the search engine results page, a significant negative effect of -1.78 was revealed (z = -2.97, p < .01). The time spent on irrelevant Web pages was not significantly related to task success (estimated effect of 0.13, z = 0.23, p = .82). As a measure of effect size, we computed Nagelkerke's R^2, which was .25, that is, about a quarter of the response variability could be explained by the component time predictors. This result pattern suggests that successful problem solvers quickly discarded the irrelevant search engine results page, whereas relevant pages meeting evaluation criteria were checked carefully. This pattern is fully compatible with the view that positive time on task effects in difficult tasks are due to a strategic allocation of cognitive resources, as already suggested by the moderation of the time on task effect by domain, task difficulty, and skill level.

Discussion

Computer-based assessment provides new possibilities to assess cognitive skills and underlying processes by measuring not only the outcome of a task but also behavioral process data that might be interpreted in terms of cognitive processes happening throughout task completion. This means that to some degree, data from computer-based assessments may be used to address research questions through means of process analysis that were previously confined to experimental research. This is of interest especially in combination with the rather large sample sizes obtained in educational assessments (compared to lab experiments). Thus, while there used to be a tradeoff— either go with small samples and deep process analysis or have large samples and test result data only— this tradeoff can be remedied to some degree by using process data from large-scale assessments.

The goal of this study was to investigate the effect of time on task on task success in reading literacy and problem solving in technology-rich environments and to test potential moderating variables. Our central hypothesis was that the relative degree of strategic versus routine cognitive processing as required by a task, as well as the test taker's acquired skill, determines the strength and direction of the time on task effect. Accordingly, our results revealed that the time on task effect was moderated by domain, task difficulty, and individual skill.

Time on Task Effects in Reading Literacy

For reading literacy, overall, a negative time on task effect was found, that is, brief times on task were associated with correct responses, and taking more time apparently was not related to greater task success. Very slow respondents thus fail on the task. This observation especially concerns easy reading tasks as shown by the negative correlation between task easiness and the task-specific time on task effect, which means that for easy tasks the time on task effect was more negative than for difficult ones. To put it simply, in very easy tasks, the correct solution was either obtained quickly or never. In contrast, for difficult reading tasks, this association got weaker and in some instances was reversed. Taking individual differences in reading skill into account, these findings were consistently extended, that is, with increasing reading skill, the time on task effect got more negative, whereas it got weaker or even positive with decreasing reading skill. Thus, for poor readers completing hard reading tasks, time on task showed a positive effect, whereas for proficient readers working on easy tasks, a very strong negative effect was found. The latter result means that the few proficient readers who did not master the easy reading tasks took more time than the majority of proficient readers who were successful. In contrast, in a group of less skilled readers, this time difference between correct and incorrect answers in the same tasks was less pronounced, as shown by the weaker negative time on task effect.

The observed result pattern that incorrect responses are associated with longer times on task has consistently been found for other untimed performance measures as well, for instance, general knowledge tasks (Ebel, 1953), matrices tasks (Hornke, 2000), figure series, number series, verbal analogy tasks (Beckmann, 2000), verbal memory tasks (Hornke, 2005), and discrimination tasks (for a review of reaction time research on this matter, see Luce, 1986). Hornke (2005) discussed how correct responses with short latencies are eye-catching. Incorrect responses in contrast may be preceded by an ongoing process of rumination and ultimately a switch to random guessing. This interpretation is consistent with our finding in that, especially for easy tasks, there is a strong negative time on task effect and also explains why, in easy reading tasks, generally skilled readers had a lower chance of getting the task correct when the response took longer. Similar effects were reported by Hornke (2000) and Beckmann (2000).

Across the cognitive operations required in reading tasks, there was no significant variation of the time on task effect. Thus, differences in the time on task effect across tasks cannot be ascribed to the presence or absence of specific cognitive operations as outlined in the PIAAC framework. In line with our findings on the dependency of the time on task effect on task difficulty, the clusters of access, integrate, and evaluate tasks are not very well distinguishable by their level of difficulty. Other task features than the cognitive operations are hence responsible for the variation of the time on task effect with task difficulty. If our cognitive interpretation of time on task effects holds, it might be worthwhile to look for task features that drive task difficulty and differential time on task effects. Identifying these features might further contribute to clarifying the PIAAC reading tasks' demands in cognitive terms and as such contribute to further advance the assessment framework. Therefore, as one future step, we intend to classify the PIAAC tasks, for instance, in terms of the transparency of the information, or the degree of complexity in making inferences (cf. OECD, 2009b). Task features such as these are not yet entirely covered by the aspects detailed by the PIAAC assessment framework.

Time on Task Effects in Problem Solving

For problem solving, overall, a significant positive time on task effect was found: Long times on task were associated with correct answers and short times on task with wrong answers. Similar to reading, the time on task effect varied significantly across tasks. For easy tasks, it was weaker and around zero, whereas for difficult tasks, it became even more positive. This means that when dealing with challenging problems, spending more time was associated with higher probability of giving a correct response. Across individuals, poor problem solvers could benefit more from spending more time on a task than strong problem solvers. Although causal interpretations are not possible, this result suggests that poor problem solvers can compensate for their lack of general skill by putting in more effort when working on a particular task, especially when this task is hard to solve. Thus, the difference in time on task between correct and incorrect solutions was greater for weak problem solvers than for strong problem solvers, which is the reverse of the finding for reading.

The results on the time on task effect for reading literacy and problem solving show that the moderating role of task difficulty and person's skill are similar for both domains, even though the overall effect is very different. The time on task effect may become similar between the two domains when considering the extreme cases in which a skilled person encounters an easy task or a less skilled person engages in a difficult task. In the first case, the resulting time on task effect is negative (even for problem solving), and in the second case, it is positive (even for reading literacy). Thus, across domains the strength and the direction of the time on task effects seem to be governed by skill and difficulty in the same way. Both high skill levels and easy tasks presumably are associated with a large proportion of cognitive component processes that are apt to automatization (in easy tasks) or in fact automatized (in skilled persons), bringing about a negative time on task effect. In contrast, low skill levels and difficult tasks presumably are associated with the need to engage in controlled and thus time-consuming cognitive activity to a large extent, bringing about a positive time on task effect.

Thus, on the one hand, problem solving and reading are conceived as involving different cognitive processes, and overall the relation of time on task to task success also clearly differs between the two domains. On the other hand, our results support the notion that combinations of tasks and persons form a continuum across the two domains ranging from automatic processing to controlled processing. Practicing a task may move a person–task combination to automatic processing. However, this is limited by the nature of the task. For instance, certain aspects of a problem solving task may become automated in skilled individuals, but not core aspects of problem solving, such as inducing rules or drawing conclusions.

Our interpretations of the time on task effect are further backed by the in-depth analysis of the time-taking behavior in the sample problem solving task "Job Search." This analysis was based on time data that was assumed to reflect different steps of task solution and presumably information processing. It revealed that only for time spent on steps that are necessarily needed to solve the tasks, that is, to visit and evaluate the target pages for multiple criteria, a positive time on task effect emerged, whereas for spending time on the noninformative search engine results page and the nontarget pages, negative or null effects were found. When spending time on the target pages, the problem solver is assumed to deal with the part of the problem space that enables one to move step by step to the knowledge state that includes the solution (Simon & Newell, 1971) or to integrate relevant information, rather than identifying various other aspects of the problem (Wirth & Leutner, 2008). Thus, this finding supported our hypothesis that the positive time on task effect in problem solving tasks reflects the need for and the benefit from devoting time to strategic and controlled cognitive processing. This interpretation suggests that task success could depend on the time spent on relevant pages (however, time on task as well as task success might also be driven by a common cause such as motivation). The negative effect of time spent on the search engine results page may indicate the strategy to select Web pages based on the limited information provided there. Although this approach could in principal be useful to filter search results, in the given task the results page did not indicate whether search criteria would be met or not. Thus, lingering on a page that could not contribute to solving the task was in fact detrimental to succeeding.

Time on Task and a Dual Processing Framework

We derived our hypotheses on differential time on task effects both between and within domains by means of applying a dual processing framework (cf. Fitts & Posner, 1967; Schneider & Chein, 2003; Schneider & Shiffrin, 1977) to reading and problem solving tasks used in the PIAAC study. The hypotheses thus derived were confirmed; hence, our results are consistent with the notion that positive time on task effects reflect the strategic allocation of cognitive resources, whereas negative time on task effects reflect the degree of automatization. Although the findings are entirely consistent with the predictions derived from such a framework and further backed by analyses on a within-task level, this interpretation has to remain somewhat speculative for the time being. The information that can be gained from large-scale computer-based assessments (although providing much more information than traditional paper-and-pencil based assessments) is still limited. Usually, the information stored in log files is ambiguous as to its interpretation in cognitive terms. In this article, we have assumed that taking more time on more difficult tasks indicates engaged cognitive processing. Other interpretations of the pattern of results are yet conceivable. For instance, it might be the case that time on task effects also reflect differences in motivation, that is, test takers not only take more time to think about a task but also think harder—resulting in a confounding between depth of processing and time taken. Related to that, Guthrie et al. (2004) considered time on task as an indicator of engagement, which means to read a text attentively, concentrating on the meaning, and with sustained cognitive effort (see also Kupiainen et al., 2014). Issues such as these can only be resolved by combining the analysis of large-scale process data with research tools allowing for an even more fine-grained analysis of cognitive processes, such as eye movements or think-aloud techniques (see Rouet & Passerault, 1999). As a consequence, we aim at corroborating our results through experimental studies that combine actual large-scale testing materials and still more fine-grained assessments of cognitive processes in the future.

Limitations

In the present study, test takers were free to adapt their speed–accuracy compromise both within and between tasks, which has consequences on the interpretation of the obtained results. As the speed level of test takers was not controlled, the obtained variation in the association between time on task and task success across tasks may be due to different task difficulties as claimed in Hypothesis 2 or due to within-person differences in the selected speed level across tasks. However, the latter explanation does not seem plausible as there is empirical evidence for the assumption of stationarity of speed when completing power tests (cf., e.g., Goldhammer & Klein Entink, 2011; Klein Entink et al., 2009). Stationarity of speed is also implied by the fixed level of accuracy which is a standard assumption in item response models (cf. van der Linden, 2007).

As we did not manipulate the speed level of test takers experimentally, we cannot conclude that the predictor time on task has any causal effect on task success, which, however, is suggested by the positive time on task effect in those tasks requiring a higher level of controlled processing. In contrast, in tasks that can be completed more automatically and for which a negative effect was revealed, time on task should rather be conceived of as an indicator of competence in addition to the task result.

As another limitation, the sample size of the present study and the number of responses per task, respectively, were quite limited for testing measurement models. Therefore, future research should aim at replicating the findings based on greater samples, for instance, from the PIAAC main study. Another important replication goal would be to investigate whether results on the time on task effect are comparable across countries.

In PIAAC the construct of problem solving in technology-rich environments was newly developed as was the measurement procedure. Thus, future research will have to provide more information about this assessment's validity and its predictive power. Moreover, the relation of problem solving in PIAAC to other problem solving measures and their theoretical underpinnings requires further clarification. There are several conceptual commonalities, for example, representing the difference between a current state and a goal state, defining a series of subgoals, and applying related nonroutine cognitive and behavioral operations to transform the given state into the targeted state, including progress monitoring. However, there are also remarkable differences. For instance, the construct of complex problem solving (cf. Funke & Frensch, 2007) assumes systems where complexity is defined by the number of elements and the relations among them. The problem solving process is comprised of the acquisition of knowledge by means of exploration and the application of the obtained knowledge. Although acquiring knowledge or information is also a key aspect of problem solving as defined in PIAAC, acquired knowledge in a complex problem solving task represents the explored system of elements and relations itself. In contrast, in PIAAC problem solving, the explored system is just the medium carrying the information that is required to solve the task. However, an unfamiliar computer environment and unknown functionality would turn the problem solving in technology-rich environments task into a complex problem solving task (for technical problem solving, see, e.g., Baumert, Evans, & Geiser, 1998). Regarding our findings on problem solving as proposed by PIAAC, future research needs to show whether the pattern of results holds true also for other conceptions of problem solving that, for instance, are anchored in cognitive theory (see, e.g., Fischer, Greiff, & Funke, 2012) or used in other large-scale assessments such as the Programme for International Student Assessment (PISA; cf. Greiff, Holt, & Funke, 2013).

Educational Implications

The present study frames the meaning of time in information-processing tasks by referring to models of skill acquisition and related individual differences. Therefore, although the analyses are based on assessment tasks, our results allow for some tentative conclusions on educational procedures in reading and problem solving instruction. Our results indicate that for learning and applying higher level cognitive skills, required component skills should be well routinized. If there is no established routine processing, for instance, when a poor reader encounters difficult reading tasks, information processing needs to rely on strategic processing as indicated by the reversed positive effect of time on task on task success. This means that for poor readers to be successful, they need to switch to compensatory behaviors, that is, reducing reading rate, looking back in text, reading aloud, and pausing, and/or compensatory strategies, that is, shifting attention to lower level requirements and rereading text, to cope with their deficits. Following Walczyk's (2000) compensatory-encoding model, "With enough time, any text can be vanquished!" (p. 565).

From an instructional point of view this means that becoming a good reader or problem solver requires the development of self-regulatory and metacognitive skills necessary to know when an effortful, controlled processing mode is to be employed. In the controlled processing mode, appropriate compensatory mechanisms can be initiated that have been learned and incorporated before. This might, for example, mean that in the face of reading comprehension difficulties, a part of a text is reread or that in problem solving, time is taken to focus attention on relevant subgoals.

As the individual time on task effect is assumed to reflect the way of processing information, it may help to further describe the individual performance level and to identify instructional needs. As suggested in Figure 4 (bottom left panel), average readers show a great variation in the time on task effect, suggesting various levels of automaticity of component skills. Moreover, the in-depth investigation of temporal patterns in highly interactive tasks such as problem solving tasks can point to deficits in the information-processing strategy (cf. Zoanetti, 2010). For instance, if, in the "Job Search" task, log file data would reveal that a problem solver spends much time both on nontarget pages and target pages, this pattern would suggest that the problem solver cannot process disconfirming information efficiently to quickly discard a nontarget page.

From an educational measurement perspective, the present study suggests that the meaning of time on task is not uniform. Thus, when collecting time information across tasks and individuals that are heterogeneous in difficulty and skill level, respectively, the role of time and its interpretation may differ. Regarding item response models including time as a regressor, van der Linden (2007) argued that time can only be interpreted uniformly as an indicator of speed if the tasks do not differ substantially in the amount of labor. In the present study where tasks differ considerably in the amount of information processing and problem solving, we take the different interpretations of time on task into account by letting its effect vary across tasks (random effect).

All in all, the analyses and results reported here illustrate the potentials that lie in exploiting time on task, or fractions of it, that become available through computer-based assessments. They do however also clarify that any process measure must be cautiously interpreted, at least by taking a closer look at the particular tasks and their demands. Regarding the two constructs studied here, reading literacy and problem solving in technology-rich environments, our study proves them to be quite different in terms of cognitive processing. Skill, task difficulty, and time on task do interact in different ways. As Wirth and Klieme (2003) have shown based on student assessment in a German national extension to PISA, problem solving tests, especially computer-based problem solving tests, add to the traditional set of literacy dimensions. In structural models, problem solving skills can be clearly distinguished from traditional abilities such as reasoning (cf. Wüstenberg, Greiff, & Funke, 2012). These structural analyses and our in-depth analyses of processing time provide evidence that problem solving skills have to be separated from traditional educational outcomes such as reading literacy. Problem solving is one of the most prominent examples of cross-curricular, nonroutine, dynamic 21st century skills that are currently aimed at as educational goals and covered in large-scale surveys. Claims that these new skills are different from traditional outcomes have mainly been supported by pragmatic or philosophical arguments. Now, we see that even in terms of cognitive processing and time allocation, there is a difference between reading literacy and problem solving skills.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3–27. doi:10.1037/0033-2909.102.1.3

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 6*, 259 –290. doi:10.1037/1076-898X.6.4.259

Anderson, J. R. (1992). Automaticity and the ACT* theory. *American Journal of Psychology, 105*, 165–180. doi:10.2307/1423026

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390 – 412. doi:10.1016/j.jml.2007.12.005

Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283–316. doi:10.1037/0096-3445.133.2.283

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using Eigen and S4 (R Package Version 0.999999 – 0) [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Baumert, J., Evans, R. H., & Geiser, H. (1998). Technical problem solving among 10-year-old students as related to science achievement, out-of-school experience, domain-specific control beliefs, and attribution patterns. *Journal of Research in Science Teaching, 35*, 987–1013. doi:10.1002/(SICI)1098-2736(199811)35:9<987::AID-TEA3>3.0.CO;2-P

Beckmann, J. F. (2000). Differentielle Latenzzeiteffekte [Differential effects of latencies]. *Diagnostica, 46*, 124 –129. doi:10.1026//0012-1924.46.3.124

Best, R., Rowe, M., Ozuru, Y., & McNamara, D. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders, 25*, 65–83. doi:10.1097/00011363-200501000-00007

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

Carlson, R. A., Khoo, B. H., Yaure, R. G., & Schneider, W. (1990). Acquisition of a problem-solving skill: Levels of organization and use of working memory. *Journal of Experimental Psychology: General, 119*, 193–214.

Carlson, R. A., Sullivan, M. A., & Schneider, W. (1989). Component fluency in a problem-solving context. *Human Factors, 31*, 489–502. doi:10.1177/001872088903100501

Carver, R. P. (1992). Reading rate: Theory, research and practical implications. *Journal of Reading, 36*, 84–95.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559. doi:10.1007/s11336-008-9092-x

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

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer. doi:10.1007/978-1-4757-3990-9

Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence, 41*, 1–10. doi:10.1016/j.intell.2012.10.003

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software, 20*, 1–18.

Ebel, R. (1953). The use of item response time measurements in the construction of educational achievement tests. *Educational and Psychological Measurement, 13*, 391– 401. doi:10.1177/001316445301300303

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving, 4*, 19 – 41. doi:10.7771/1932-6246.1118

Fitts, P. M., & Posner, M. I. (1967). *Learning and skilled performance in human performance*. Belmont, CA: Brooks/Cole.

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

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective—10 years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25– 47). New York, NY: Erlbaum.

Gelman, A. (2005). Analysis of variance? Why it is more important than ever. *Annals of Statistics, 33*, 1–53. doi:10.1214/009053604000001048

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence, 39*, 108 –119. doi:10.1016/j.intell.2011.02.001

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment*. Advance online publication. doi:10.1027/1015-5759/a000153

Gräsel, C., Fischer, F., & Mandl, H. (2000). The use of additional information in problem-oriented learning environments. *Learning Environments Research, 3*, 287–305. doi:10.1023/A:1011421732004

Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on problem solving in educational assessment: Analytical, interactive, and collaborative problem solving. *Journal of Problem Solving, 5*, 71–91. doi:10.7771/1932-6246.1153

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational settings—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*, 364 –379. doi:10.1037/a0031856

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., . . . Tonks, S. (2004). Increasing reading comprehension and engagement through Concept-Oriented Reading Instruction. *Journal of Educational Psychology, 96*, 403– 423. doi:10.1037/0022-0663.96.3.403

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica, 21*, 175–189.

Hornke, L. F. (2005). Response time in computer-aided testing: A “verbal memory” test for routes and maps. *Psychology Science, 47*, 280 –293.

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

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*, 21– 48. doi:10.1007/s11336-008-9075-y

Klieme, E. (2004). Assessment of cross-curricular problem-solving competencies. In J. H. Moskowitz & M. Stephens (Eds.), Comparing learning outcomes: *International assessments and education policy* (pp. 81– 107). London, England: Routledge.

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19*, 513–526. doi:10.1016/j.learninstruc.2008.10.002

Kupiainen, S., Vainikainen, M.-P., Marjanen, J., & Hautamäki, J. (2014). The role of time on task in computer-based assessment of cross-curricular skills. *Journal of Educational Psychology, 106*, 627–638. doi:10.1037/a0035507

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100*, 150 –161. doi:10.1037/0022-0663.100.1.150

Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization*. Oxford, England: Oxford University Press.

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York, NY: Freeman.

Mayer, R. E. (1994). Problem solving, teaching and testing for. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., Vol. 8, pp. 4728 – 4731). Oxford, England: Pergamon Press.

McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly, 44*, 218 –253. doi:10.1598/RRQ.44.3.1

Morrell, C. H., Pearson, J. D., & Brant, L. J. (1997). Linear transformations of linear mixed-effects models. *American Statistician, 51*, 338 –343. doi:10.1080/00031305.1997.10474409

Naumann, J., Richter, T., Christmann, U., & Groeben, N. (2008). Working memory capacity and reading skill moderate the effectiveness of strategy training in learning from hypertext. *Learning and Individual Differences, 18*, 197–213. doi:10.1016/j.lindif.2007.08.007

Naumann, J., Richter, T., Flender, J., Christmann, U., & Groeben, N. (2007). Signaling in expository hypertexts compensates for deficits in reading skill. *Journal of Educational Psychology, 99*, 791– 807. doi:10.1037/0022-0663.99.4.791

Neubauer, A. C. (1990). Speed of information processing in the Hick paradigm and response latencies in a psychometric intelligence test. *Personality and Individual Differences, 11*, 147–152. doi:10.1016/0191-8869(90)90007-E

Nicolson, R. I., & Fawcett, A. J. (1994). Reaction times and dyslexia. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 47*(A), 29 – 48. doi:10.1080/14640749408401142

Organisation for Economic Co-Operation and Development. (2009a). *PIAAC literacy: A conceptual framework* (OECD Education Working Paper No. 34). Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2009b). *PIAAC problem solving in technology-rich environments: A conceptual framework* (OECD Education Working Paper No. 36). Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2011). *PISA 2009 results: Vol. VI. Digital technologies and performance*. Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. doi:10.1787/9789264204256-en

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357–383. doi:10.1080/10888430701530730

Puntambekar, S., & Stylianou, A. (2005). Designing navigation support in hypertext systems based on navigation patterns. *Instructional Science, 33*, 451– 481. doi:10.1007/s11251-005-1276-5

R Core Team. (2012). R: A language and environment for statistical computing [Computer software]. Retrieved from http://www.R-project.org/

Richter, T., Isberner, M.-B., Naumann, J., & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern [Process-based measurement of reading skills in primary school children]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 26*, 313–331. doi:10.1024/1010-0652/a000079

Richter, T., Isberner, M.-B., Naumann, J., & Neeb, Y. (2013). Lexical quality and reading comprehension in primary school children. *Scientific Studies of Reading*. Advance online publication. doi:10.1080/10888438.2013.764879

Richter, T., Naumann, J., Brunner, M., & Christmann, U. (2005). Strategische Verarbeitung beim Lernen mit Text und Hypertext [Strategic processing in learning from text and hypertext]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 19*, 5–22. doi:10.1024/1010-0652.19.12.5

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York, NY: Springer. doi:10.1007/978-1-4757-2691-6_11

Rouet, J.-F., & Passerault, J.-M. (1999). Analyzing learner-hypermedia interaction: An overview of on-line methods. *Instructional Science, 27*, 201–219. doi:10.1007/BF00897319

Scheuermann, F., & Björnsson, J. (2009). *The transition to computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.

Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education, 54*, 627– 650. doi:10.1007/s11159-008-9105-0

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science, 27*, 525–559. doi:10.1016/S0364-0213(03)00011-9

Schneider, W., & Fisk, A. D. (1983). Attentional theory and mechanisms for skilled performance. In R. A. Magill (Ed.), *Memory and control of action* (pp. 119 –143). New York, NY: North-Holland. doi:10.1016/S0166-4115(08)61989-5

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1– 66. doi:10.1037/0033-295X.84.1.1

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190. doi:10.1037/0033-295X.84.2.127

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

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. S. Haladyna (Eds.), *Handbook of test development* (pp. 329 –347). Mahwah, NJ: Erlbaum.

Sullivan, S., Gnesdilow, D., & Puntambekar, S. (2011). Navigation behaviors and strategies used by middle school students to learn from a science hypertext. *Journal of Educational Multimedia and Hypermedia, 20*, 387– 423.

van den Broek, P. W., & Espin, C. A. (2012). Connecting cognitive theory and assessment: Measuring individual differences in reading comprehension. *School Psychology Review, 41*, 315–325.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308. doi:10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*, 247–272. doi:10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195–210. doi:10.1177/01466219922031329

Walczyk, J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly, 35*, 554 –566. doi:10.1598/RRQ.35.4.7

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics*. Acta Psychologica, 41*, 67– 85. doi:10.1016/0001-6918(77)90012-9

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice, 10*, 329 –345. doi:10.1080/0969594032000148172

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods*. Zeitschrift für Psychologie/Journal of Psychology, 216*, 102–110. doi:10.1027/0044-3409.216.2.102

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

Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. Australasian Journal of Educational Technology, 26, 585– 606.