Eichmann, Beate; Goldhammer, Frank; Greiff, Samuel; Brandhuber, Liene; Naumann, Johannes

Using process data to explain group differences in complex problem solving

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Journal of educational psychology 112 (2020) 8, S. 1546-1562

This is a post-peer-review, pre-copyedit version of an article published in *The Journal of Educational Psychology*. The final authenticated version is available online at: http://dx.doi.org/10.1037/edu0000446

Using Process Data to Explain Group Differences in Complex Problem Solving

Beate Eichmann[a], Frank Goldhammer[a], Samuel Greiff[b], Liene Brandhuber[c], Johannes Naumann[d]

a: DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB)

b: University of Luxembourg

c: Goethe University Frankfurt

d: University of Wuppertal

Author Note

Corresponding author: Beate Eichmann, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB) Rostocker Straße 6, 60323 Frankfurt am Main, Germany, Tel: +49 69 24708-863, beate.eichmann@dipf.de

This research was funded by the German Federal Ministry of Education and Research (grant numbers: 01LSA1504A and 01LSA1504B) and by a project funded by the Fonds National de la Recherche Luxembourg (The Training of Complex Problem Solving; "TRIOPS").

Parts of this research were presented at the NCME 2018 Annual Meeting and at the GEBF 2018 Annual Meeting.

Abstract

In large scale assessments, performance differences across different groups are regularly found. These group differences (e.g. gender differences) are often relevant for educational policy decisions and measures. However, the formation of these group differences usually remains unclear. We propose an approach for investigating this formation by considering behavioral process measures as mediating variables between group membership and performance on the 2012 PISA complex problem solving items. We found that across all investigated countries interactive behavior can fully explain gender differences in CPS, but cannot explain differences between students with and without a migration background. However, in some countries these results differ from the cross-country results. Our results indicate that process measures derived from log data are useful for further investigating and explaining performance differences between girls and boys and students with and without migration background.

Keywords: Complex problem solving, exploration, log data, computer-based assessment, PISA

Educational Impact and Implications Statement

The study suggests that the higher performance of boys compared to girls in complex problem solving seems to stem from gender-specific interaction with the problem space, while performance differences by migration status cannot be explained by behavioral differences. Specifically, the amount of exploration behavior seems to have a huge impact on complex problem solving performance. The study demonstrates how the formation of performance differences between different groups of students can be explained.

Using Process Data to Explain Group Differences in Complex Problem Solving

In educational settings, performance differences between demographic groups are regularly observed (OECD, 2011, OECD, 2014a; Prenzel et al., 2004). Group variables that are associated with performance differences include students' gender or migration background. For instance, in large scale assessments like the Programme for International Student Assessment (PISA), girls regularly outperform boys in the domain of reading, and boys regularly outperform girls in mathematics (OECD, 2011, OECD, 2014a). These performance differences raise questions of equity and fairness in the educational systems in which they occur, which sometimes leads to changes in educational policy. For instance, after the results of PISA 2000 were published, national educational standards were established in Germany to reduce performance differences associated with social background (Neumann, Fischer, & Kauertz, 2010). To understand the causes of and eventually minimize these performance differences, it is important to investigate their underlying mechanisms. For instance, in the domain of literacy, Artelt, Naumann, and Schneider (2010) found that girls outperform boys in reading partly due to a better command of metacognitive strategies.

Another domain where performance differences between demographic groups arise regularly is problem solving. For example, in PISA 2012, several group-level performance differences in problem solving were observed across OECD countries (OECD, 2014b). One factor related to problem solving performance was gender; boys scored significantly higher than girls. The difference was 0.07 standard deviations on average across countries. Another factor related to problem solving ability was migration background; students without a migration background scored significantly higher than students with a migration background. On average, this difference amounted to 0.32 standard deviations (OECD, 2014b).

Complex problem solving

A problem is defined as a situation in which a person wants to achieve a goal but no obvious solution is available (Mayer & Wittrock, 2006). Problems can further be divided into the subdomains of analytical (static) and complex (dynamic) problems. Complex problem solving (CPS) is a particularly important competency in today's society and can be regarded as a 21st century skill since it is required in many situations in everyday life (Binkley et al., 2012). In contrast to analytical problems, complex problems change dynamically (over time or as a result of interaction), and not all information required to solve the problem is present at the outset (Frensch & Funke, 1995). Therefore, complex problems require the problem solver to interact with the problem to obtain necessary information, while analytical problems can be solved just by thinking. Due to its interactive nature, CPS is usually measured via computer-based assessment. In computer-based assessment, log data is available that can give insight into the task solution process (Greiff, Wüstenberg, & Avvisati, 2015; Kroehne & Goldhammer, 2018). This means that differences between groups in terms of cognitive and behavioral task engagement can be analyzed as one possible source of differences in task success, and thus in estimated ability. In the present study, we investigate whether performance differences in CPS between demographic groups can be explained by behavioral differences extracted from log data.

Problem solving process: Interaction and exploration behavior

What behavioral processes might account for individual success, and thus possibly also group differences, in CPS? Some predictions can be made on the basis of the literature on error management training. Error management training refers to an active learning approach in which trainees are encouraged to explore a system, even if this means they might commit errors. The core elements of error management training are experimentation and exploration without providing much guidance. Trainees are supposed to commit errors to learn how to handle the system and errors within the system (see Frese & Keith, 2015 for an overview). Error management training is often used to train people to use software systems. A number of studies have shown that error management training is superior to error-avoidant training, especially in novel situations that require knowledge transfer, and thus in situations which could also be characterized as problems (e.g. Bell & Kozlowski, 2008; Keith & Frese, 2005; Keith, Richter, & Naumann, 2010; see Keith & Frese, 2008, for a meta-analysis). Dormann and Frese (1994) investigated the effect of exploration in error management training on performance in a complex technology-based environment. Participants received either error-avoidant training or error training within a statistics program before their performance was assessed. While the subjects in the error-avoidant group received a structured tutorial that prescribed every step necessary to accomplish specific tasks with the statistics program, the error training group only received a set of basic commands and no further introduction. Participants were selected to have basic statistical experience and basic experience with similar software. The study's authors defined exploration during training as interactions with the environment that had not been previously introduced. They found higher performance in the error training group than in the error-avoidant group and a positive effect of exploration in the training session on subsequent performance. The superior performance of the error training group was shown to be attributable to trainees in the error training engaging in more comprehensive exploration, and thereby acquiring a deeper understanding and becoming familiar with more possible states of the system (Dormann & Frese, 1994). However, Dormann and Frese (1994) state that exploration should not be confused with trial-and-error, since exploration should be guided by hypotheses based on a mental model of the system. Kapur (2008) also argues that exploration leads to higher performance in problem solving. He showed that exploration led to a better representation of and to higher knowledge about the problem. Bell and Kozlowski (2008) also reported a positive effect of exploration in training on subsequent performance in a complex computer-based simulation. They compared the performance of university students instructed to use exploration in a training phase with those who received a structured step-by-step tutorial. Students in the exploration group were instructed to explore the system to discover suitable methods and strategies for handling it, while students in the other group received detailed step-by-step instructions on how to accomplish goals and were asked to follow these instructions. Both groups were provided with the same list of training objectives to achieve. The students in the exploration group exhibited higher performance in the follow-up assessment than students who received the structured tutorial. Bell and Kozlowski (2008) argue that exploration provides learners with control over their learning process, which in turn activates their metacognition (e.g. planning, monitoring and revising behavior). As both Bell and Kozlowski (2008), and Keith and Frese (2005) argue, metacognitive processes activated by exploration enhance learning and transfer. For example, Bell and Kozlowski (2008) argue that self-evaluation activities are positively related to participants' strategic knowledge, intrinsic motivation and self-efficacy. Drawing upon these perspectives, we argue that explorative behavior is a crucial prerequisite for successful CPS, since the cognitive processes mentioned above (e.g. activation of metacognition, problem representation, hypotheses testing) also apply to CPS.

Besides exploration, the overall amount of interaction with a problem scenario might also predict CPS performance. Interaction with a problem means that the problem solver takes any observable action. In a computer-based scenario this could be, for example, a mouse click or a keystroke. Naumann, Goldhammer, Rölke, and Stelter (2014) investigated effects of interaction on problem solving success in technology-based environments using data from the Programme for the International Assessment of Adult Competencies (PIAAC) (see OECD., 2013b). They found that the number of interactions had a quadratic relation with problem solving performance. The inverse u-shape had its optimum at 1.5 standard deviations above average; thus, it seems that students who refrain from interaction particularly struggle. Naumann et al. (2014) argue that low computer-related self-efficacy or high computer-related anxiety might be the reasons for persons to behave passively and therefore be less successful in solving technology-based problems. Since in PIAAC the participants' basic computer skills were assessed and the computer-based problem solving test was only administered to those participants who showed sufficient computer skills, the unfamiliarity with computers is unlikely to be the reason for the passive behavior of some participants. On the basis of their results, Naumann et al. (2014) argue that in problem-solving in technology-based environments, acting too cautiously might pose a greater threat to performance than acting too boldly. This interpretation is in line with the literature on error management training, where training approaches that encourage "risky" behavior when learning – i.e. willingness to commit an error, and learn from it – is recommended (Bell & Kozlowski, 2008; Dormann & Frese, 1994).

In the present study, we distinguish between interaction behavior and exploration behavior, with exploration representing a subset of overall interactions. More specifically, we define interaction as a student's engagement with a problem irrespective of what exactly they do. Further, we define exploration as a student's interaction with aspects of a problem that provide information about the problem situation but do not directly contribute to its solution. For instance, if the problem requires the student to buy a daily ticket for the subway, clicking on the button for individual trips on the ticket machine does not provide an instant solution and therefore would be defined as exploration behavior, as it might help the student build a better representation of the problem space. Repeated interaction with solution-relevant aspects of a problem is also regarded as exploration as long as it is not essential for solving the problem. Hence, if the student in the example above would go back and press the button for a daily ticket again after having already pressed it, the second button press would be regarded as exploration (because the student could have already solved the problem after the first button press). Conversely, the first press on the button for a daily ticket is classified as an instance of interaction, but not exploration (since this button press is directly goal-oriented).

Explaining group differences in CPS

Both theory and empirical results point to the importance of the amount of interaction and exploration in CPS on the individual level (Bell & Kozlowski, 2008; Dormann & Frese, 1994; Frensch & Funke, 1995; Naumann et al., 2014). Building upon this perspective, in the present research, we ask whether the amount of interaction and exploration can also account for CPS performance differences *between groups* where they exist. Note that whether group-level differences in *complex* problem solving exist at all is not a trivial question. PISA 2012 demonstrated differences between boys and girls, and between students with and without a migration background in overall problem solving, that is, in a measure of problem solving that entailed both complex *and* analytical problem solving. On this basis, it is plausible to assume that similar performance differences would emerge if complex problem solving would be specifically examined.

If indeed performance differences in CPS exist between girls and boys, these might be traced back to behavioral differences associated with gender. Wittmann and Hattrup (2004) found behavioral differences between females and males that might account for gender differences in

CPS performance. They argue that males outperformed females because males engage in more risky behavior, which in turn provoked more dramatic changes in the CPS scenario, and in consequence, provided them with more information and more learning opportunities about the system. These findings are in line with the results of a meta-analysis by Cross, Copping, and Campbell (2011) indicating that males in general exhibited more risk-taking behavior than females. Wüstenberg, Greiff, Molnár, and Funke (2014) also found higher CPS performance among male students compared to females. They observed that boys applied the VOTAT strategy (VOTAT=vary one thing at a time) associated with high CPS performance more often. Wüstenberg et al. (2014) found similar differences in the application of the strategy between girls and boys in different countries. Sonnleitner, Brunner, Keller, and Martin (2014) investigated behavioral differences in CPS between students with and without a migration background. They found that students with a migration background exhibited more exploration behavior than their peers without a migration background. However, students without a migration background exhibited higher CPS performance. Sonnleitner et al. (2014) explained this result by arguing that students with a migration background might have had difficulties transferring the generated information into declarative knowledge. Martin, Liem, Mok, and Xu (2012) identified socio-economic status, language background, age of migration, gender, and age as factors that are relevant to immigrant students' problem solving performance. From the available evidence, there seems to be good grounds to assume that differences in interaction and exploration might indeed account for performance differences in CPS between boys and girls. In contrast, there is a much weaker basis for this assumption regarding performance differences in CPS between students with and without migration.

Thus, the first aim of the present study was to investigate whether the performance differences between boys and girls and between students with and without a migration background that PISA 2012 found for overall problem solving would also be present in the CPS subdomain. If this would be the case, the second aim was to investigate whether these effects are mediated through behavioral patterns while solving complex problems (OECD, 2014b), specifically interaction and exploration.

Hypotheses

Performance differences between groups

On the basis of previous results demonstrating performance differences in CPS between boys and girls (Wüstenberg et al., 2014) and students with and without a migration background (Sonnleitner et al., 2014), we expect to find performance differences between these groups as well:

Hypothesis 1a: Boys exhibit higher CPS performance than girls.

Hypothesis 1b: Students without a migration background exhibit higher CPS performance than students with a migration background.

Effects of behavioral indicators on CPS performance

As Frensch and Funke (1995) claim, CPS requires interacting with the problem. Moreover, the amount of interaction can predict performance in technology-based problem solving (Naumann et al., 2014). The work of Bell and Kozlowski (2008) and Dormann and Frese (1994) showed that exploration behavior has a positive impact on performance in complex environments that have a problem-like character. Therefore, we hypothesize that the amount of interaction and exploration is also related to CPS performance:

Hypothesis 2a: The number of interactions is positively related to CPS performance.

Hypothesis 2b: The number of exploration steps is positively related to CPS performance.

Behavioral differences between groups

As mentioned above, Wittmann and Hattrup (2004) and Wüstenberg et al. (2014) found behavioral differences between females and males when solving complex problems. The former observed more risky behavior in males, while the latter observed that males apply the VOTAT strategy more often. In both studies, boys exhibited higher performance and more behavior associated with high performance. Both risky behavior and the VOTAT strategy might be associated with exploration. When exploring, students risk committing errors. The VOTAT strategy requires students to explore a problem scenario to obtain information required for solving the problem. Therefore, we expect boys to engage in more exploration than girls, and since exploration is part of interaction, we also expect boys to engage in more interaction than girls.

The findings by Sonnleitner et al. (2014) indicate that students with a migration background engage in more exploration behavior than their fellow students without a migration background. Therefore, we assume that students with a migration background engage in more exploration and thus also (since exploration is part of interaction) more interaction than students without a migration background.

Hypothesis 3a: Boys exhibit more interactive and explorative behavior than girls.

Hypothesis 3b: Students with a migration background exhibit more interactive and explorative behavior than students without a migration background.

Mediation of performance differences

From previous theory and findings, there is ample reason to hypothesize that performance differences in complex problem solving between boys and girls might be mediated through different styles of engagement, and specifically through interaction and exploration. Boys tend to be more prone to engage in “risky” behavior (Wittmann & Hattrup, 2004), and thus to engage in exploration, which is a useful strategy in complex problem solving:

Hypothesis 4a: The effect of gender on complex problem solving performance is mediated through exploration behavior.

Since, in addition to exploration, previous research has shown that interaction in the sense of both directly task-related behavior as well as exploratory behavior is mostly positively related to CPS performance (Naumann et al., 2014), interaction might also be suspected to mediate the effects of gender:

Hypothesis 4b: The effect of gender on complex problem solving performance is mediated through interaction behavior.

In contrast to gender, the role of behavioral differences between students with and without a migration background in bringing about performance differences between the two groups is much less clear. For example, while students with a migration background lag behind their peers without a migration background, they tend to do *more* exploration (Sonnleitner et al., 2014), behavior that is considered here to be beneficial for complex problem solving. Thus, while differences in the behavioral indicators considered here might account for CPS performance differences related to migration background, other variables like language proficiency might be even more crucial (Martin et al., 2012). For these reasons, we refrain from specifying a hypothesis regarding the mediating role of exploration and interaction with respect to the effects of migration background. Instead, we treat the question of whether the effects of migration background on CPS performance are mediated through exploration and interaction as an exploratory research question.

Method

Sample

We used log data generated during the computer-based assessment of problem solving in PISA 2012. We only used data from the 44 countries that participated in the computer-based assessment of problem solving. However, of those countries, we had to exclude Cyprus because data on gender and migration status were not available. We also decided to exclude Korea since not all of our models converged with the Korean dataset. Therefore, the overall sample size was further reduced to N=81,039 students from 42 countries (50.15% female, 12.22% with a migration background). The number of participating students, the percentage of female students and the percentage of students with a migration background in each country are shown in Appendix A.

Instruments

The CPS assessment in PISA 2012 consisted of 27 computer-based items, which were organized into 16 units alongside analytical problem solving items. The conceptual distinction between complex (called interactive in PISA) and analytical items (called static in PISA) was made by the OECD (2013a). Their criterion to distinguish between the two item types was the disclosure of information about the problem. In contrast to static problems, in interactive problems not all information was disclosed at the outset to the problem solver. Therefore, we argue that this definition of interactivity matches the definition of CPS by Frensch and Funke (1995) we follow in this article. Each unit was comprised of two to three items with similar stimulus material. Students worked on one or two out of four different problem solving clusters, which included four units each. The item order within units was always the same. After finishing an item, students could not return to it.

The items were embedded in everyday contexts, and they were designed to control for prior knowledge by applying very heterogeneous contexts (OECD, 2013a). Moreover, prior knowledge was not sufficient to solve a problem solving item in PISA 2012. Examples of these everyday contexts are controlling room temperature and humidity using a climate control panel and buying train tickets from a ticket machine. Prior knowledge about ticket machines might help to solve this item, however the item still requires CPS activities to be solved. In principal, prior knowledge could have been avoided by constructing abstract or artificial contexts for a task. However, problem solving items without a meaningful context might lack external validity to everyday problems and, therefore, lack relevance. Moreover, in PISA 2012 Differential Item Functioning analyses were carried out to make sure that all items worked equally well for students of different gender, of different language proficiency or from different countries. In this procedure problematic items were revised or excluded (OECD, 2014c).

The computer-based CPS items required only basic computer skills like clicking on virtual buttons and sliders, dragging and dropping, operating simulated machines, exploring simulated environments, and manipulating variables. The response formats included simple and complex multiple-choice items that were answered by clicking radio buttons (26% of the items), items in which selections had to be made from pull-down menus (7% of the items), items that required parts of diagrams to be drawn (26% of the items), items that required establishing a certain state by clicking buttons (37% of the items) and text boxes (11% of the items) (OECD, 2013a). Note that some items contained more than one response format. Before the assessment, a tutorial was administered so that students could practice the required skills. The tutorial was offered to all students. However, students could choose whether they wanted to work on the tutorial or skip it. For a detailed explanation of the item characteristics, see OECD (2013a).

Gender and migration status were assessed via a student questionnaire. Students were asked whether they or their parents were born abroad. As recommended by OECD (2014c), we defined a migration background as having two parents who were born outside the country of assessment. We did not differentiate between first and second generation migration background.

Scoring

We used the available log data to score students' responses. For CPS performance, the response coding suggested by OECD (2015b) was used: The responses were coded as either correct (1) or incorrect (0). Therefore, our CPS scoring indicates whether or not a problem was completely solved. The following example should illustrate the scoring: The problem to be solved is to buy the cheapest ticket from a ticket machine that would meet certain criteria (e.g. a ticket for several trips with the city subway including a student discount). Since the problem solvers are not familiar with the ticket machine they cannot know what fares and ticket types are available and how to choose between them. This item was coded as being correctly solved if the problem solver bought the ticket that met all the criteria specified in the task. If the problem solver bought a different ticket or did not buy a ticket at all, the item would be coded as not correctly solved. To rule out any confounds between our process measures and CPS performance, we did not use the original PISA scoring rules that included scores for students' behavior during the task. Out of the 27 CPS items, two items were excluded because the respective log files included clicks on elements with ambiguous IDs. Another two items had to be excluded because the responses included free text input, which was not recorded in the log data, so the correctness of the response could not be inferred. Four items were excluded because the total number of user interactions (which is relevant for our analyses) was externally restricted. Another three items were excluded because interaction with the item was automatically stopped as soon as the correct system state was reached, also limiting the possibilities for exploration. After these exclusions, a total of 16 items remained out of the 27 CPS items (for item examples see OECD, 2013a).

We extracted two process measures from the log data: (1) the overall number of interactions and (2) the number of exploration steps. For the overall number of interactions, we counted all click events that occurred for each item completion sequence. To determine the number of exploration steps, we defined the shortest possible click pattern that would lead to a correct solution for each item. For some items, more than one shortest click pattern was defined, since there were multiple equally short ways to solve the item. We then calculated the Levenshtein distance (LD) between these shortest click patterns and the students' actual click patterns. The LD is a measure of the difference between two sequences. It counts the number of insertions, deletions and substitutions of sequence elements necessary to transform one sequence into the other (Navarro, 2001). For items with more than one shortest click pattern, the LDs between every shortest pattern and the students' patterns were calculated, meaning that every student had several LDs for these items. Then, we took the minimum LD for each student for further analysis to obtain a conservative estimate of the amount of exploration. When a student's click pattern is identical to the shortest possible pattern, the LD equals zero. Hence, an LD of zero indicates that no exploration took place. The LD increases with an increasing number of clicks. However, the LD also increases with a decreasing number of clicks for click patterns shorter than the shortest possible solution pattern. To get a valid measure for the number of exploration steps, we therefore adjusted the LD for those shorter click patterns by subtracting the difference between the length of the shortest possible pattern and the length of the actual click pattern. This gave us an indicator of the number of exploration steps that counted every interaction beyond the “minimal” item solution, i.e. an item solution process that does not entail any kind of exploratory behavior. We expect the distribution of the number of exploration steps to be right-skewed, since it is a mere count of students’ exploration steps. Table 1 provides an example of how the number of exploration steps was computed. In the example the task has two possible shortest click patterns to come to the correct solution (pattern 1: A-B-C, pattern 2: A-D-E-F). Student 1 has a click pattern longer than these shortest patterns (A-D-E-F-G-H). Therefore, no adjustment is necessary. Student 1’s number of exploration steps equals the minimum LD between his pattern and the two shortest patterns (which is two). However, student 2 hast a click pattern shorter than the shortest possible patterns (A-D). Therefore, the LD needs to be adjusted. The adjustment is accomplished by subtracting the difference between the length of student 2’s pattern and the respective shortest patterns (i.e. -1 for pattern 1 and -2 for pattern 2). After the adjustment, the minimum LD equals the number of student 2’s exploration steps (which is zero).

Procedure

PISA 2012 covered mathematics, reading, science, problem solving, and financial literacy. Problem solving, mathematics and reading were assessed via computer-based assessment. Students were administered two computer-based clusters of which none, one, or both were problem solving clusters. Since the problem solving clusters were randomly assigned to the students, parameters should be comparable across items. Also the pairs of clusters overlapped meaning that all possible combinations of item clusters were administered. The computer-based assessment took place after the completion of the paper-based PISA items. Students had 20 minutes to complete each cluster (OECD, 2014b).

Data Preparation

Data preparation was performed separately for every country. First, the two process indicators were inspected for outliers. An outlier was defined as a value three standard deviations above/below the respective average. The outliers were replaced with the value exactly three standard deviations above/below the average, as suggested by Goldhammer et al. (2014). We identified outliers corresponding to 1.08% of the values for the number of interactions and 2.71% of the values for the amount of exploration. Because both indicators showed high skewness, we log-transformed the indicators after adding +1 to every value to enable logarithmic transformation of zeros. The log transformation was necessary to normalize the data in order to perform a maximum likelihood estimation. By log transforming we changed the metric of our data leading to a numeric reduction of differences in the upper value range.

Data Analysis

We estimated four mediation models for each country. In each model, the independent variable was either gender or migration background, and the mediating variable was either interaction or exploration. The criterion variable in each case was CPS performance, defined as correctly solving the respective problems. Both gender and migration background were dichotomous (male=0, female=1; no migration background=0, migration background=1). The mediator variables were modeled as latent variables on the person level, aggregating the respective process indicator across all CPS items. The item intercepts were estimated freely to account for different demands for interactions across items. The dependent variable (CPS performance) was also modeled as a latent variable on the person level, aggregating the scores across all CPS items. . In all our models we controlled for reading ability as a possible confound. Since reading ability is required to understand and answer PISA's CPS tasks properly, students

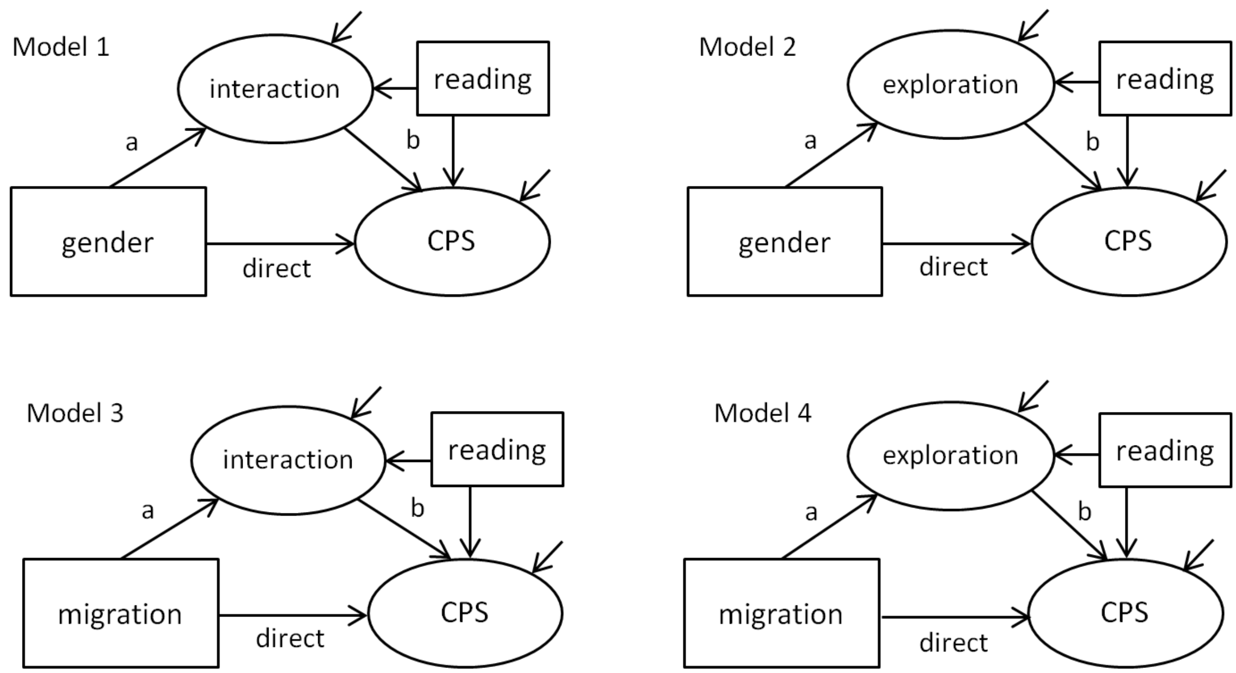

Figure 1: The four resulting models.

with migration background might experience higher difficulty due to lower language ability. Therefore, Martin et al. (2012) suggest controlling for reading ability when investigating CPS among students with migration background. But also between girls and boys differences in reading ability are regularly observed which might decrease boys' performance in CPS tasks (OECD, 2014a). Therefore, we decided to control for reading ability in all our analyses. We used weighted likelihood estimators based on the items from the PISA 2012 paper-based reading assessment to control for reading ability in the mediator variables and in CPS performance. The resulting models are shown in Figure 1. To account for PISA's complex sampling design, the final PISA student weights were used and school was included as a cluster variable. We used Mplus 7.4 for the mediation analyses (Muthén & Muthén, 1998-2015) and the R package MplusAutomation for executing the country-specific calculations (Hallquist & Wiley, 2018; R Core Team, 2016). The effect size κ^2 was calculated for the mediation effect of every country using the R package MBESS (Kelley & Lai, 2010). Following procedures suggested by

Naumann (2015, 2019), we employed random-effects meta-analysis (see Hedges & Vevea, 1998) to integrate country-wise results, providing both an estimate of a fixed effect that is the same across countries, as well as an estimate of each effects variance across countries. We used the R package metafor (Viechtbauer, 2010) to summarize the mediation model parameters across countries, treating each country as a “study” in a meta-analysis. As a measure of between country variation, we report the square root of estimated between-country variance τ, the Q-statistic for testing τ^2 for heterogeneity, and its p-value. The effect size κ^2 was aggregated using its median.

Results

Table 2 shows the aggregated standardized model results and effect sizes. The model results and effect sizes by country are listed in Appendix B, Appendix C, Appendix D, and Appendix E. An overview of the estimated models is given in Figure 1.

Performance differences between groups (Hypothesis 1a and 1b)

We found a significant total effect of gender on CPS performance, indicating that boys outperformed girls in both our models incorporating gender (Model 1: total effect=-0.28, $p<.001$; Model 2: total effect=-0.28, $p<.001$). This supports Hypothesis 1a. Also as expected, there was a total effect of migration background on CPS performance indicating that students without a migration background outperformed students with a migration background (Model 3: total effect =-0.16, $p=.001$; Model 4: total effect =-0.12, $p=.003$). Therefore, Hypothesis 1b was supported as well. However, both observed performance differences varied significantly between countries (Model 1: τ(total effect)=0.09, $p<.001$; Model 2: τ(total effect)=0.09, $p=.002$; Model 3: τ(total effect)=0.22, $p<.001$; Model 4: τ(total effect)=0.19, $p<.001$). Although the aforementioned performance differences were found in most countries, in 1 country the effect of gender and in 11 countries the effect of migration background on performance was reversed. For example, in the United Arab Emirates girls outperformed boys and in Poland immigrant students outperformed students without a migration background (see Appendix B, Appendix C, Appendix D, Appendix E).

Effects of behavior indicators on CPS performance (Hypothesis 2a and 2b)

The results from Model 1 and Model 3 showed that the number of interactions was positively related to CPS performance (Model 1: $b=0.71$, $p<.001$; Model 3: $b=0.74$, $p<.001$). The number of exploration steps was also positively related to CPS performance (Model 2: $b=0.44$, $p<.001$; Model 4: $b=0.44$, $p<.001$). Thus, Hypothesis 2a and Hypothesis 2b were supported. The effect of the number of interactions exhibited significant between-country variation, while the effect of exploration did not (Model 1: $\tau(b)=0.06$, $p<.001$; Model 3: $\tau(b)=0.06$, $p<.001$; Model 2: $\tau(b)=0.03$, $p=.355$; Model 4: $\tau(b)=0.05$, $p=.052$). However, in all countries both effects pointed in the same direction (see Appendix B, Appendix C, Appendix D, Appendix E).

Behavioral differences between groups (Hypothesis 3a and 3b)

Gender. The results of Model 1 and Model 2 showed that gender was negatively related to the number of interactions and the number of exploration steps, indicating that boys exhibited more interactions and more exploration than girls (Model 1: $a=-0.27$, $p<.001$; Model 2: $a=-0.57$, $p<.001$), providing support to Hypothesis 3a. The relation between gender and exploration was stronger than the relation between gender and interactions. We found significant between-country variations for both effects (Model 1: $\tau(a)=0.12$, $p<.001$; Model 2: $\tau(a)=0.09$, $p=.002$). The estimates pointed in the same direction in all countries for the number of exploration steps and in all countries except one for the number of interactions (see Appendix B, Appendix C).

However, in the United Arab Emirates the relation between gender and number of interactions was reversed.

Migration Background. We did not find a significant association between migration background and number of interactions (Model 3: a=-0.09, p=.094). However, we found significant between-country variation (Model 3: τ(a)=0.27, p<.001). We did not find a relation between migration background and the amount of exploration (Model 4: a=0.02, p=.665). Once again, we found significant between-country variation of this effect (Model 4: τ(a)=0.25, p<.001). Therefore, Hypothesis 3b is not supported.

Mediation of performance differences (Hypothesis 4a and b and exploratory research question)

Gender. The negative effect of gender on CPS performance was mediated by the number of interactions. The indirect effect was negative and significant (Model 1: indirect effect=-0.19, p<.001) and exhibited significant between-country variation (Model 1: τ(indirect effect)=0.07, p<.001), meaning that boys presumably exhibited stronger CPS performance due to performing a larger number of interactions. The direct effect was also significant (Model 1: direct effect =-0.08, p<.001) and did not exhibit between-country variation (Model 1: τ(direct effect)=0.04, p=.241). This meant that while overall boys outperformed girls in CPS, *conditional on a specific number of interactions*, the performance difference was clearly smaller. The median κ^2 was 0.13, which reflects a medium to large sized mediation effect (Cohen, 1988), supporting Hypothesis 4b.

The amount of exploration seemed to be an even stronger mediator. The indirect effect was negative (Model 2: indirect effect =-0.23, p=.077) and did not vary between countries (Model 2: τ(indirect effect)=0.03, p=.382). This meant that the better performance of boys as compared to girls was not only related to the larger amount of interaction overall, but specifically related to a larger amount of exploration. The direct effect was not significant (Model 2: direct effect =-0.03, p=.083). This meant that after controlling for exploration, girls and boys showed equal CPS performance. The direct effect also showed no between-country variation (Model 2: τ(direct effect)=0.05, p=.292). The median κ^2 was 0.17, which reflects a medium to large sized effect according to Cohen's conventions (1988), and indicates a larger effect than in Model 1.

Migration Background. The results from Model 3 indicated no mediation of the effect of migration background on CPS performance by the number of interactions. The indirect effect was not significant (Model 3: indirect effect=-0.07, p=.077). It did, however, vary between countries (Model 3: τ(indirect effect)=0.19, p<.001). The direct effect was not significant either, (Model 3: direct effect=-0.07, p=.146; total effect=-0.16, p=.001) and also showed between-country variation (Model 3: τ(direct effect)=0.27, p<.001). The direct effect not being significant meant that after controlling for the number of interactions, and reading skill, students without a migration background did not outperform students with a migration background.

The effect of migration background on CPS performance was not mediated by the amount of exploration. The indirect effect was not significant (Model 4: indirect effect=0.02, p=.235). It did, however, vary between countries (Model 4: τ(indirect effect)=0.08, p<.001). The direct effect was of similar size as the total effect and significant (Model 4: direct effect=-0.15, p<.001; total effect=-0.12, p=.003). The direct effect also varied between countries (Model 4: τ(direct effect)=0.15, p<.001). Therefore, we conclude that performance differences in CPS between students with and without a migration background cannot be explained by the amount of interaction or the amount of exploration.

Discussion

In the present study, we used process measures to explain performance differences in CPS between groups. We used data from over 81,000 students from 42 countries. Therefore, those results that did not show between-country variation can be generalized across many different cultures. We first investigated performance differences in CPS between girls and boys and between students with and without a migration background. Second, we investigated how interactive and explorative processes are related to CPS performance. Third, we investigated whether students of different gender or migration status differed in their interactive and explorative behavior. Finally, we tested whether the observed performance differences can be explained by differences in interactive or explorative behavior. In the following section, we will discuss our results.

Performance differences between groups

Our results indicate that boys exhibit higher performance in CPS than girls. In PISA 2012, boys exhibited higher problem solving performance than girls (OECD, 2014b). Our finding extends the results of PISA 2012 to the subdomain CPS. The results also show that students with a migration background exhibited lower performance than students without a migration background. Again, this extends PISA 2012 findings in which students with a migration background exhibited lower performance in problem solving (OECD, 2014b). Notably, both group differences showed between-country variation. In some countries, the observed group differences were not present at all or were even reversed. For instance, in the United Arab Emirates, girls exhibited higher CPS performance than boys, and in Singapore, students with a migration background exhibited higher performance than students without a migration background. Therefore, when referring to specific countries, the specific findings and patterns of those countries should be considered as well. We will discuss possible causes for the observed performance differences in the section “Mediation of performance differences”.

Effects of behavior indicators on CPS performance

Both of our process measures (number of interactions and number of exploration steps) were positively related to CPS performance. This finding is in line with the results of Naumann et al. (2014), who showed that less interaction is detrimental for problem solving performance in technology-based environments. We also extended the findings of Bell and Kozlowski (2008) and Dormann and Frese (1994) concerning exploration in error training to CPS. However, the relation between interaction and CPS was stronger than the relation between exploration and CPS. This indicates that goal-directed behavior is more important than exploration for solving a complex problem. This finding is not surprising, since a complex problem cannot be solved without goal-directed behavior; meanwhile, exploration should facilitate the execution of goal-directed behavior but is not immediately necessary to solve a problem. These effects also showed between-country variation. However, both effects were positive in all countries, making a strong argument for a general positive relation between interaction/exploration and CPS performance.

Behavioral differences between groups

As expected, we found behavioral differences between boys and girls. Boys exhibited more interactive behavior and also more exploration. These results extend the findings of Wittmann and Hattrup (2004), who found that boys exhibit more risky behavior, and Wüstenberg et al. (2014), who showed that boys used the VOTAT strategy more often. Both risky behavior as described by Wittmann and Hattrup (2004) and the VOTAT strategy as described by Wüstenberg et al. (2014) can be regarded as special cases of exploration behavior. The risky behavior in the study by Wittmann and Hattrup (2004) involved the use of extreme values in a simulated problem scenario. On the one hand, the use of these extreme input values carried the risk of causing "catastrophic" effects in the simulation. On the other hand, causing more dramatic changes in the simulation by choosing extreme input values provided more information about the structure of the problem than moderate values would have produced. Therefore, this behavior can be regarded as exploration. The VOTAT strategy described by Wüstenberg et al. (2014) can also be seen as a special case of exploration behavior. The VOTAT strategy is applicable to items in the linear structural equation systems framework (Greiff & Funke, 2009). It implies that every input variable should be manipulated separately to find out which relations between input and output variables exist. Students generate more knowledge about the problem structure by applying this strategy more frequently (Wüstenberg et al., 2014), which is why it might also be regarded as exploration behavior. Therefore, our findings can be regarded as a generalization of these previous results. Although we found between-country variation, the estimates in all countries for the number of exploration steps and in the majority of countries for the number of interactions pointed in the same direction. An exception was again the United Arab Emirates, where we found that females engaged in more interactions than males. These differences in the relation of gender and behavior between countries could be related to the respective culture and its predominant gender norms. Culture-specific gender norms could influence the behavior of girls and boys in different ways, including differences in education for girls and boys both at school and at home.

We found no relation between migration background and number of interactions, indicating that students with a migration background interacted as much with the problems as students without a migration background. This finding does not support the results of Sonnleitner et al. (2014), who found that immigrant students in Luxembourg interacted more with the problems than their peers without a migration background. However, the immigrant population in Luxembourg may have particular characteristics that do not occur in other countries. We will further discuss this issue in the section “Mediation of performance differences”. We did not find any effect of migration background on the amount of exploration. Therefore, Sonnleitner et al.’s argumentation (2014) that immigrant students do not have a deficit in exploration behavior but have difficulties utilizing the generated information can be applied to our results. We also found significant between-country variation and a substantial number of countries with reversed estimates for both effects. Therefore, these effects should be interpreted in a country-conditional manner.

Mediation of performance differences

Our results showed that the performance difference between boys and girls was mediated by both process measures, with amount of exploration being a stronger mediator than the number of interactions, despite the fact that the number of interactions was the stronger predictor of CPS performance. In other words, a lack of interaction and even more so a lack of exploration prevent girls from exhibiting equally high CPS performance as boys. One reason why girls exhibit less interaction and exploration might be that girls are encouraged less often to engage in this kind of behavior than boys. For example, Cherney and London (2006) argue that play with different kinds of toys may foster the development of different cognitive abilities in girls and boys. They found that boys between 5 and 13 years of age preferred toys that encourage manipulation, construction, and exploration, while girls in the same age group preferred toys that encourage the development of verbal skills. Leaper and Friedman (2007) argue that children start even earlier to develop gender-related cognitions. They state that three-year-old children are already aware of their own gender-group membership which becomes part of their social identity. Between 3 and

6 years of age, children begin to form stereotypes about gender-specific activities. Therefore, girls might not be motivated or may not even come up with the idea of exhibiting exploration behavior and therefore exhibit lower CPS performance. Miller (1987) also argues that girls are discouraged to solve problems by certain socialization practices such as the discouragement of active play and the restriction of exploration through parents, teachers, peers, the media, and cultural institutions. Therefore, these processes could be promising starting points for improving girls' CPS performance. This difference in the socialization of girls and boys may differ between cultures, since we found between-country variation of the relation of gender and interaction/exploration behavior. Another reason why girls might be less motivated to engage in interaction and exploration behavior could be lower self-efficacy in computer-based environments. For example, one finding of PISA 2012 was that in most countries boys are exposed to computers much earlier than girls (OECD, 2015b). The OECD (2015b) argues that restricting girls' access to computers might lower their self-efficacy in computer-based tasks. The OECD (2015c) also found that on average girls spend less of their leisure time engaging with computers and that they less often have career ambitions in the field of computing and engineering than boys. Spending less time with computers girls might indeed develop a lower computer-related self-efficacy than boys. Again, these relations may differ between countries, which is reflected by the between-country variation of the relation between gender and behavior we found. However, it seems implausible that differences in computer skills caused the observed CPS performance differences since only basic computer skills were needed to solve the items and a tutorial was used to make sure that all participants were able to operate the computer-based testing environment. Moreover, the OECD (2016) found in their PIAAC study that basic computer skills are quite balanced across females and males and that poor basic computer skills

are rarely found among young adults. Punter, Meelissen, and Glas (2017) did also not find gender differences in applying technical functionality using data from the International Computer and Information Literacy Study (ICILS). Moreover, Greiff, Kretzschmar, Müller, Spinath, and Martin (2014) found only weak to moderate relations between CPS ability and computer skills in several studies. For these reasons, we think it unlikely that the CPS performance differences we observed are a function of computer skills differing between genders. Although not lacking the required computer skills, assessing CPS using computer-based items could have discouraged girls to some extent (due to lower computer-related self-efficacy), leading them to exhibit less interactive behavior and thus lower performance than boys. Again, these effects show between-country variation, so when referring only to a single country its specific effect should be considered. For instance, in Estonia no mediation of the gender differences by the number of interactions was observed. If indeed gender performance differences result from gender-specific childhood experiences regarding the encouragement of behavior and access to technology, the differences in CPS between countries might be partly a result of cultural differences in gender norms and socialization. This socialization could also lead to lower technology-related self-efficacy or higher anxiety among girls which in turn leads more passive behavior as Naumann et al. (2014) argue. However, interaction and exploration can also be subject to educational interventions in the respective countries. In the United Arab Emirates in which we found girls to show more interactions the educational system might promote this kind of behavior (see Appendix B).

We found that the performance difference between students with and without a migration background was neither mediated by the number of interactions nor by the amount of exploration. Sonnleitner et al. (2014) argue that immigrant students engage in more exploration than students without a migration background. However, they state that immigrant students have difficulties transferring the generated information into declarative knowledge and therefore are outperformed in CPS by students without a migration background. Our results support the view that immigrant students exhibit exploration behavior equal to non-immigrant students. Although exhibiting exploration behavior they might not be able to process the required information. One possible reason why students with migration background might not profit from exploration behavior (see Table 2) might be a low proficiency in the test language. However, since we controlled for reading ability in our models, neither a low language ability nor a lack of exploration behavior seem to be the primary cause for the low CPS performance of immigrant students. Sonnleitner et al. (2014) investigated the relation between CPS and migration background. They found that the lower performance of immigrant students could be explained by students with migration background being much more often enrolled in lower academic tracks. Also Greiff et al. (2013) found a strong relation between students' CPS ability and their academic achievement.

Theoretically, the country-specific mechanisms leading to lower CPS performance among students with a migration background could be an effect of the composition of the migrant population in each country. For example, in countries in which most immigrants received little education, these deficits in education could be the main reason for immigrants' underachievement (e.g. countries mostly recruiting workers for rather simple jobs from abroad, leading to immigration of low educated people). However, in countries in which most immigrants are highly educated, there could be other reasons for performance differences (e.g. countries recruiting mostly highly educated people from abroad). Another factor that might be related to the CPS performance of students with a migration background is their socio-economic status. In PISA 2012, socio-economic status (SES) was related to many competencies, including problem solving (OECD, 2014b). Therefore, the socio-economic composition of the immigrant population in each country could also affect their CPS performance. In this case, it would not be appropriate to generalize this effect across countries. The SES could also be related to students' access to technology, since low SES households might not be able to afford computers or laptops. Therefore, lower computer abilities among low SES students might be a confounding variable to the computer-based CPS assessment. However, as stated before the tasks only required very rudimental computer skills, and a tutorial was offered, so that every student should have been able to solve the tasks as far as operating the technological interface is concerned. Moreover, Greiff et al. (2014) found only weak to moderate relations between CPS and computer skills. Future research however might further address these questions by examining the socio-economic status and level of education of migrants in different countries. Interaction effects between gender and migration that might vary between countries would also be conceivable and might be addressed in future research. Moreover, not only the country of assessment but also the countries where migrants originate from might play a role in this regard.

Limitations

Since our data refers only to fifteen-year-old students, the generalizability of our results to different age groups is limited. Like all studies that use PISA data we cannot rule out possible confounding variables like computer skills in our analyses. Furthermore, we had to exclude the data from Korea since not all models converged for this country's dataset. Another limitation of our study is the fact that we had to exclude several items. For future research, it would be preferable to have more complete and unambiguous log data to avoid item drop-outs. We also had to exclude some items due to a restriction in exploration. The exclusion of these items might have led to an over- or underrepresentation of certain item characteristics in the resulting item pool. It should also be kept in mind that we did not use any experimental manipulation. Therefore, our results are of a correlational nature and should be further confirmed with experimental designs in future research. Moreover, we only used two somewhat arbitrary behavioral indicators to represent interactive and explorative behavior. Especially with regard to exploration, different operationalizations would also be possible. Since we defined all interactions that were not necessary for item completion as exploration steps, our definition was rather broad. However, this broad definition was necessary to align with the heterogeneous items in the PISA assessment. Another limitation of our findings is that we cannot explain the variation in the observed effects between countries. Future research should try to explain this variation by examining country characteristics, for instance differences in the composition of immigrant groups. But also differences in the respective curricula could lead to country-specific effects. Especially, CPS being part of a country's curriculum could heavily influence the results. Furthermore, a country's error culture could influence the extent to which students are willing to explore which might lead to errors. This between-country variation also limits the generalizability of our results making it necessary to take into account the context of the countries of interest when interpreting the results.

The differences between the effects of gender and migration background between countries might also be related to cultural differences in gender roles or differences in the migration population. For example Naumann, Elson, and Rauch (2016, April) found that in digital reading, a domain with close ties to complex problem solving (Brand-Gruwel, Wopereis, & Walraven, 2009), there were stark differences between economies in the effect of migration background on task-adaptive navigation. These effects were strong in European, but absent or weak in Oceanic and Chinese economies, where Chinese students, who generally perform very well abroad (OECD, 2015a), make a large part of the immigrant population. A more detailed analysis of this issue however must be left to future research, using data sets where explicit information on the countries immigrant students migrated from is available.

In a similar vein, future research might look at interactive effects between gender and migration background, conditional on the culture from where immigrant students came, on CPS behavior and performance. It might well be the case that disadvantages for girls and for students with migration background overlap and reinforce each other.

Moreover, not differentiating between first and second generation migration background might have had weakened the effects found with respect to migration background. The OECD (2014) reported an even lower problem solving performance of students who were born outside the country of assessment than students who were born in the country of assessment (but whose parents were born abroad). Therefore, in future research these groups could be investigated separately to reveal differential effects.

Conclusion

Our results indicated that the performance difference in CPS between boys and girls can be explained by interaction and exploration behavior. Since exploration is part of overall interactive behavior and exploration more strongly mediates the gender effect, explorative behavior might just be the crucial factor causing performance differences between girls and boys in CPS. As soon as interaction or exploration was taken into account, girls' performance was equal to boys' performance. Therefore, girls' lower performance in CPS might be due to a lower readiness to explore. On the other hand, the performance difference between students with and without a migration background cannot be explained by explorative or interactive behavior.

Notably, in some countries, students with a migration background exhibited more interactions than students without a migration background, while in other countries, this effect was reversed. Nevertheless, students with a migration background exhibited lower performance in most countries. Therefore, behavioral differences do not seem to be the primary cause for the lower CPS performance of students with a migration background.

In our study, we could show that measures derived from log data can serve as mediating variables to explain performance differences between groups and thus shed light on the mechanisms behind these performance differences. We could show that gender differences in CPS performance seem to stem from gender-specific behavior, while differences by migration status seem to have other causes. Therefore, besides showing that log data can be used to predict performance, we also showed that performance differences between groups can be explained by behavioral differences as recorded in log data. This means that future studies using computer-based assessment data could report not only group differences in performance, but also possible causes of these differences.

References

Artelt, C., Naumann, J., & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel,. . . P. Stanat (Eds.), *PISA 2009: Bilanz nach einem Jahrzehnt* (pp. 73–112). Münster: Waxmann.

Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, *93*, 296–316. https://doi.org/10.1037/0021-9010.93.2.296

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining Twenty-First Century Skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 17–66). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2324-5_2

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

Cherney, I. D., & London, K. (2006). Gender-linked Differences in the Toys, Television Shows, Computer Games, and Outdoor Activities of 5- to 13-year-old Children. *Sex Roles*, *54*, 717–726. https://doi.org/10.1007/s11199-006-9037-8

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: a meta-analysis. *Psychological Bulletin*, *137*, 97–130. https://doi.org/10.1037/a0021591

Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction*, *6*, 365–372. https://doi.org/10.1080/10447319409526101

Frensch, P. A., & Funke, J. (1995). Definitions, Traditions, and a General Framework for Understanding Complex Problem Solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 3–22). Hillsdale, NJ: L. Erlbaum Associates.

Frese, M., & Keith, N. (2015). Action errors, error management, and learning in organizations. *Annual Review of Psychology*, *66*, 661–687. https://doi.org/10.1146/annurev-psych-010814-015205

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

Greiff, S., & Funke, J. (2009). Measuring Complex Problem Solving: The MicroDYN approach. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 157–163).

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, *106*, 666–680. https://doi.org/10.1037/a0035426

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

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

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–18. https://doi.org/10.1080/10705511.2017.1402334

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*, 379–424. https://doi.org/10.1080/07370000802212669

Keith, N., & Frese, M. (2005). Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology*, *90*, 677–691. https://doi.org/10.1037/0021-9010.90.4.677

Keith, N., & Frese, M. (2008). Effectiveness of error management training: a meta-analysis. *Journal of Applied Psychology*, *93*, 59–69. https://doi.org/10.1037/0021-9010.93.1.59

Keith, N., Richter, T., & Naumann, J. (2010). Active/Exploratory training promotes transfer even in learners with low motivation and cognitive ability. *Applied Psychology*, *59*, 97–123. https://doi.org/10.1111/j.1464-0597.2009.00417.x

Kelley, K., & Lai, K. (2010). MBESS: MBESS. R package. Retrieved from https://CRAN.R-project.org/package=MBESS

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log-data from technology-based assessments? - A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*, 527–563.

Leaper, C., & Friedman, C. K. (2007). The Socialization of Gender. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of Socialization: Theory and Research* (pp. 561–587). New York: Guilford Publications.

Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: Multination findings from the Programme for International Student Assessment (PISA). *Journal of Educational Psychology*, *104*, 1054–1073. https://doi.org/10.1037/a0029152

Mayer, R. E., & Wittrock, M. C. (2006). Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–304). Mahwah, N.J: Erlbaum.

Miller, C. L. (1987). Qualitative Differences Among Gender-Stereotyped Toys: Implications for Cognitive and Social Development in Girls and Boys. *Sex Roles*, *16*, 473–487.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, *53*, 263–277. https://doi.org/10.1016/j.chb.2015.06.051

Naumann, J. (2019). The Skilled, the Knowledgeable, and the Motivated: Investigating the Strategic Allocation of Time on Task in a Computer-Based Assessment. *Frontiers in psychology*, *10*.

Naumann, J., Elson, M., & Rauch, D. (2016, April). *Explaining performance gaps in digital reading between native and immigrant students through group-specific navigation behavior*. AERA Annual Meeting, Washington, D.C.

Naumann, J., Goldhammer, F., Rölke, H., & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen. *Zeitschrift für Pädagogische Psychologie*, *28*, 193–203. https://doi.org/10.1024/1010-0652/a000134

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*, 31–88. https://doi.org/10.1145/375360.375365

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in germany. *International Journal of Science and Mathematics Education*, *8*, 545–563. https://doi.org/10.1007/s10763-010-9206-7

OECD. (2011). *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)* (1. Aufl.). *PISA 2009 Results*: OECD. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=655759

OECD. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. *PISA*. Paris: OECD.

OECD. (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. Paris: OECD Publishing. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=1683156

OECD. (2014b). *PISA 2012 Results: Creative Problem Solving: Students' skills in tackling real-life problems* (Volume V). Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264208070-en

OECD. (2014c). *PISA 2012: Technical Report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2015a). *Immigrant Students at School: Easing the Journey towards Integration*. *OECD Reviews of Migrant Education*. Paris: OECD Publishing.

OECD. (2015b). *Students, computers and learning: Making the connection*. Paris: OECD Publishing.

OECD. (2015c). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. *PISA*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264229945-en

OECD. (2016). *Skills matter: Further results from the survey of adult skills*. *OECD skills studies*. Paris: OECD Publishing.

OECD. (2013b). *OECD skills outlook 2013*. Paris: OECD Publishing.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115. https://doi.org/10.1037/a0022658

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M.,. . . Schiefele, U. (Eds.). (2004). *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland : Ergebnisse des zweiten internationalen Vergleichs*. Münster, New York: Waxmann.

Punter, R. A., Meelissen, M. R. M., & Glas, C. A. W. (2017). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal*, *16*, 762–780.

R Core Team. (2016). *R: A language and environment for statistical computing: R Foundation for Statistical Computing*. Vienna. Retrieved from https://www.R-project.org/

Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, *106*, 681–695. https://doi.org/10.1037/a0035506

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, *21*, 393–409. https://doi.org/10.1002/sres.653

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, *29*, 18–29. https://doi.org/10.1016/j.lindif.2013.10.006

Tables

Table 1: Example calculation of the number of exploratory steps for an item with two possible shortest click patterns for two different students. LD is only adjusted for student 2, since this student's click pattern is shorter than the shortest possible patterns for the task. After the adjustment the smaller number of the two LDs is chosen.

	Click pattern student 1: A-D-E-F-G-H	Click pattern student 2: A-D
Levenshtein dictance without adjustment:		
Shortest pattern 1: A-B-C	5	2
Shortest pattern 2: A-D-E-F	2	2
Levenshtein dictance after adjustment:		
Shortest pattern 1: A-B-C	5	1
Shortest pattern 2: A-D-E-F	2	0
Number of exploration steps	2	0

Table 2: Aggregated model estimates and effect sizes

	parameter	estimate	SE	z	p	τ	Q(41)	p
Model 1	a	-0.27	0.02	-12.32	<.001	0.12	135.18	<.001
	b	0.71	0.01	62.86	<.001	0.06	100.18	<.001
	total	-0.28	0.02	-14.24	<.001	0.09	99.40	<.001
	direct	-0.08	0.01	-6.42	<.001	0.04	46.99	.241
	indirect	-0.19	0.01	-13.25	<.001	0.07	98.26	<.001
	κ^2	0.13						
Model 2	a	-0.57	0.02	-27.14	<.001	0.09	71.88	.002
	b	0.44	0.02	30.32	<.001	0.03	43.76	.355
	total	-0.28	0.02	-14.44	<.001	0.09	98.84	.002
	direct	-0.03	0.02	-1.73	.083	0.05	45.44	.292
	indirect	-0.23	0.01	-19.68	<.001	0.03	43.08	.382
	κ^2	0.17						
Model 3	a	-0.09	0.05	-1.67	.094	0.27	182.00	<.001
	b	0.74	0.01	60.95	<.001	0.06	106.92	<.001
	total	-0.16	0.05	-3.45	.001	0.22	154.70	<.001
	direct	-0.07	0.05	-1.45	.146	0.27	199.50	<.001
	indirect	-0.07	0.04	-1.77	.077	0.19	172.56	<.001
	κ^2	0.05						
Model 4	a	0.02	0.05	0.43	.665	0.25	141.55	<.001
	b	0.44	0.02	29.76	<.001	0.05	56.72	.052
	total	-0.12	0.04	-2.97	.003	0.19	122.96	<.001
	direct	-0.15	0.04	-4.33	<.001	0.15	79.51	<.001
	indirect	0.02	0.02	1.19	.235	0.08	87.96	<.001
	κ^2	0.03						

Note. Model 1 and 2: independent variable is gender. Model 3 and 4: independent variable is migration background. Model 1 and 3: mediator is number of interactions. Model 2 and 4: mediator is amount of exploration. Parameter a represents the effect of the independent variable on the mediator. Parameter b represents the effect of the mediator on the dependent variable.

Appendix A

Appendix A. Sample size, percentage of females and percentage of students with a migration background by country.

Country	n	% female	% with migration background
Australia	5608	49.34	19.08
Austria	1328	50.98	16.64
Belgium	2145	49.84	14.08
Brazil	1455	50.03	1.03
Bulgaria	2138	48.97	0.65
Canada	4584	50.22	18.08
Chinese Taipei	1483	52.06	0.74
Colombia	2286	54.07	0.22
Croatia	1923	50.81	12.01
Czech Republic	3076	50.26	3.38
Denmark	1948	52.67	24.64
Estonia	1363	51.14	7.85
Finland	3531	48.17	12.29
France	1344	51.93	13.54
Germany	1350	48.52	9.93
Hong Kong-China	1323	46.64	31.67
Hungary	1300	51.92	1.77
Ireland	1188	51.43	9.43
Israel	1341	56.67	17.52
Italy	1370	45.11	7.88
Japan	3011	48.12	0.33
Macao-China	1564	49.94	59.40
Malaysia	1927	51.32	1.50
Montenegro	1845	52.57	6.23
Netherlands	1752	48.92	10.62
Norway	1237	48.42	9.46
Poland	1227	50.29	0.16
Portugal	1444	50.21	7.55
Russian Federation	1537	49.06	9.82
Serbia	1775	51.21	8.62
Shanghai-China	1203	51.29	0.91
Singapore	1392	49.28	16.95
Slovak Republic	1463	45.80	0.68
Slovenia	2064	45.06	9.45
Spain	2703	50.13	8.95
Sweden	1256	52.15	15.29
Switzerland	1575	52.25	1.40
Turkey	1995	48.12	0.80
United Arab Emirates	3246	50.89	52.34

United Kingdom	1456	52.95	13.26
United States of America	1271	50.83	18.96
Uruguay	2012	52.09	0.60

Appendix B

Appendix B: Model 1 results and effect size κ² by country with gender as predictor and number of interactions as mediator

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ²
Australia	-0.18	0.05	0.66	0.03	-0.06	0.05	-0.12	0.03	-0.18	0.04	0.06
Austria	-0.39	0.10	0.66	0.05	-0.22	0.09	-0.26	0.07	-0.48	0.08	0.19
Belgium	-0.19	0.06	0.72	0.03	-0.07	0.06	-0.14	0.05	-0.21	0.06	0.09
Brazil	-0.28	0.10	0.75	0.07	-0.19	0.10	-0.21	0.08	-0.40	0.10	0.13
Bulgaria	-0.18	0.06	0.70	0.03	-0.18	0.06	-0.13	0.04	-0.31	0.06	0.05
Canada	-0.24	0.07	0.61	0.04	-0.04	0.06	-0.14	0.04	-0.18	0.06	0.08
Chinese Taipei	-0.37	0.09	0.62	0.06	-0.09	0.09	-0.23	0.06	-0.32	0.08	0.18
Colombia	-0.26	0.08	0.71	0.06	-0.26	0.10	-0.18	0.06	-0.45	0.09	0.13
Croatia	-0.33	0.06	0.76	0.03	-0.19	0.05	-0.25	0.05	-0.44	0.06	0.14
Czech Republic	-0.40	0.07	0.65	0.04	0.03	0.05	-0.26	0.04	-0.23	0.05	0.20
Denmark	-0.21	0.09	0.72	0.07	-0.30	0.10	-0.15	0.06	-0.45	0.10	0.05
Estonia	-0.16	0.10	0.73	0.06	-0.17	0.10	-0.12	0.07	-0.29	0.10	0.00
Finland	-0.35	0.06	0.70	0.04	0.04	0.06	-0.25	0.05	-0.21	0.06	0.10
France	-0.24	0.08	0.69	0.05	-0.11	0.07	-0.17	0.06	-0.28	0.08	0.09
Germany	-0.14	0.08	0.74	0.04	-0.15	0.07	-0.10	0.06	-0.25	0.08	0.00
Hong Kong-China	-0.30	0.08	0.74	0.05	-0.10	0.08	-0.22	0.06	-0.32	0.09	0.17
Hungary	-0.27	0.08	0.69	0.06	-0.01	0.09	-0.19	0.06	-0.19	0.09	0.08
Ireland	-0.03	0.11	0.68	0.06	-0.16	0.11	-0.02	0.08	-0.18	0.11	0.03
Israel	-0.25	0.09	0.78	0.04	-0.16	0.07	-0.20	0.07	-0.36	0.09	0.11
Italy	-0.38	0.14	0.83	0.04	-0.04	0.11	-0.32	0.12	-0.35	0.11	0.30
Japan	-0.38	0.06	0.67	0.04	-0.07	0.07	-0.25	0.05	-0.32	0.07	0.20
Macao-China	-0.48	0.10	0.66	0.07	-0.01	0.12	-0.32	0.08	-0.33	0.12	0.21
Malaysia	-0.19	0.07	0.67	0.04	0.00	0.06	-0.13	0.04	-0.13	0.06	0.06
Montenegro	-0.15	0.08	0.80	0.03	0.03	0.09	-0.12	0.06	-0.09	0.11	0.06
Netherlands	-0.04	0.09	0.80	0.04	0.00	0.06	-0.03	0.07	-0.03	0.08	0.07
Norway	-0.22	0.10	0.71	0.06	-0.03	0.09	-0.16	0.07	-0.19	0.09	0.08
Poland	-0.34	0.10	0.74	0.05	0.04	0.10	-0.25	0.08	-0.22	0.08	0.20
Portugal	-0.30	0.10	0.76	0.05	-0.14	0.09	-0.23	0.08	-0.36	0.10	0.18
Russian Federation	-0.24	0.09	0.66	0.05	-0.08	0.08	-0.16	0.06	-0.24	0.09	0.07
Serbia	-0.35	0.07	0.70	0.06	-0.14	0.06	-0.25	0.05	-0.39	0.07	0.14
Shanghai-China	-0.53	0.09	0.55	0.06	-0.01	0.10	-0.29	0.06	-0.31	0.09	0.20
Singapore	-0.47	0.08	0.56	0.06	0.01	0.09	-0.27	0.05	-0.26	0.08	0.18
Slovak Republic	-0.38	0.09	0.60	0.05	-0.17	0.08	-0.23	0.06	-0.40	0.08	0.16
Slovenia	-0.06	0.09	0.81	0.05	-0.12	0.08	-0.05	0.07	-0.17	0.09	0.07

Spain	-0.23	0.10	0.86	0.04	-0.05	0.10	-0.20	0.09	-0.25	0.10	0.18
Sweden	-0.20	0.09	0.80	0.04	-0.16	0.08	-0.16	0.07	-0.32	0.09	0.02
Switzerland	-0.45	0.11	0.76	0.06	-0.10	0.11	-0.34	0.09	-0.45	0.09	0.26
Turkey	-0.59	0.06	0.62	0.04	-0.10	0.08	-0.36	0.05	-0.46	0.07	0.22
United Arab Emirates	0.12	0.07	0.73	0.04	-0.02	0.06	0.08	0.05	0.06	0.07	0.20
United Kingdom	-0.35	0.09	0.62	0.08	-0.20	0.12	-0.22	0.06	-0.42	0.12	0.16
United States of America	-0.35	0.13	0.57	0.07	-0.02	0.10	-0.20	0.08	-0.22	0.10	0.13
Uruguay	-0.20	0.06	0.78	0.03	-0.03	0.06	-0.16	0.05	-0.18	0.07	0.09

Note. a: effect of gender on interaction. b: effect of interaction on CPS.

Appendix C

Appendix C: Model 2 results and effect size κ^2 by country with gender as predictor and amount of exploration as mediator.

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ^2
Australia	-0.56	0.06	0.45	0.07	0.07	0.06	-0.25	0.04	-0.18	0.05	0.18
Austria	-0.67	0.11	0.49	0.10	-0.15	0.12	-0.33	0.09	-0.48	0.08	0.25
Belgium	-0.53	0.10	0.52	0.07	0.05	0.08	-0.28	0.06	-0.23	0.06	0.19
Brazil	-0.28	0.24	0.36	0.10	-0.30	0.14	-0.10	0.09	-0.40	0.10	0.06
Bulgaria	-0.52	0.09	0.47	0.10	-0.06	0.09	-0.25	0.07	-0.31	0.07	0.13
Canada	-0.51	0.09	0.33	0.10	-0.02	0.08	-0.17	0.05	-0.18	0.06	0.11
Chinese Taipei	-0.64	0.10	0.56	0.09	0.03	0.11	-0.36	0.10	-0.33	0.08	0.28
Colombia	-0.50	0.15	0.70	0.13	-0.14	0.16	-0.35	0.13	-0.49	0.09	0.27
Croatia	-0.53	0.08	0.50	0.06	-0.18	0.08	-0.26	0.05	-0.44	0.06	0.17
Czech Republic	-0.65	0.07	0.42	0.07	0.04	0.07	-0.28	0.05	-0.23	0.05	0.21
Denmark	-0.48	0.20	0.60	0.65	-0.17	0.40	-0.29	0.38	-0.46	0.10	0.18
Estonia	-0.42	0.12	0.53	0.08	-0.08	0.11	-0.22	0.07	-0.30	0.10	0.11
Finland	-0.67	0.07	0.28	0.11	-0.01	0.11	-0.19	0.08	-0.20	0.06	0.12
France	-0.50	0.10	0.39	0.08	-0.10	0.09	-0.19	0.05	-0.30	0.08	0.12
Germany	-0.50	0.14	0.45	0.13	-0.05	0.12	-0.23	0.10	-0.28	0.08	0.16
Hong Kong-China	-0.57	0.10	0.50	0.09	-0.04	0.10	-0.28	0.08	-0.32	0.10	0.19
Hungary	-0.47	0.16	0.38	0.12	-0.02	0.11	-0.18	0.09	-0.20	0.10	0.13
Ireland	-0.51	0.17	0.33	0.18	-0.01	0.14	-0.17	0.10	-0.18	0.11	0.11
Israel	-0.39	0.13	0.52	0.10	-0.15	0.10	-0.21	0.07	-0.36	0.08	0.12
Italy	-0.80	0.22	0.48	0.18	0.01	0.22	-0.38	0.21	-0.37	0.11	0.24
Japan	-0.61	0.06	0.45	0.06	-0.05	0.07	-0.28	0.05	-0.32	0.07	0.17
Macao-China	-0.70	0.11	0.44	0.08	-0.01	0.12	-0.31	0.07	-0.33	0.12	0.18
Malaysia	-0.53	0.08	0.46	0.07	0.11	0.08	-0.24	0.05	-0.13	0.07	0.17
Montenegro	-0.41	0.08	0.49	0.08	0.10	0.11	-0.20	0.05	-0.10	0.11	0.10
Netherlands	-0.23	0.14	0.41	0.16	0.06	0.08	-0.09	0.04	-0.04	0.07	0.07
Norway	-0.72	0.10	0.36	0.11	0.09	0.12	-0.26	0.09	-0.17	0.09	0.19
Poland	-1.01	0.14	0.57	0.20	0.36	0.26	-0.58	0.25	-0.22	0.08	0.39
Portugal	-0.48	0.10	0.56	0.09	-0.11	0.11	-0.27	0.07	-0.38	0.10	0.18
Russian Federation	-0.69	0.10	0.26	0.08	-0.06	0.11	-0.18	0.06	-0.24	0.09	0.11
Serbia	-0.52	0.08	0.44	0.07	-0.15	0.08	-0.23	0.05	-0.38	0.07	0.14
Shanghai-China	-0.75	0.14	0.35	0.12	-0.05	0.13	-0.26	0.10	-0.31	0.09	0.16
Singapore	-0.65	0.09	0.37	0.09	0.00	0.11	-0.24	0.07	-0.25	0.08	0.16
Slovak Republic	-0.82	0.10	0.32	0.10	-0.13	0.12	-0.26	0.08	-0.39	0.08	0.20
Slovenia	-0.50	0.14	0.79	0.15	0.20	0.13	-0.40	0.12	-0.20	0.09	0.32
Spain	-0.55	0.12	0.53	0.12	0.04	0.12	-0.29	0.09	-0.25	0.11	0.18

Sweden	-0.55	0.13	0.42	0.13	-0.13	0.11	-0.23	0.08	-0.36	0.09	0.14
Switzerland	-0.69	0.12	0.49	0.08	-0.12	0.13	-0.34	0.09	-0.45	0.09	0.19
Turkey	-0.67	0.08	0.33	0.08	-0.24	0.09	-0.22	0.06	-0.46	0.07	0.15
United Arab Emirates	-0.18	0.10	0.56	0.06	0.15	0.08	-0.10	0.06	0.05	0.07	0.00
United Kingdom	-0.74	0.08	0.48	0.09	-0.08	0.13	-0.35	0.08	-0.43	0.12	0.20
United States of America	-0.70	0.13	0.28	0.08	-0.02	0.10	-0.20	0.07	-0.22	0.10	0.12
Uruguay	-0.43	0.11	0.33	0.08	-0.05	0.08	-0.14	0.06	-0.19	0.07	0.09

Note. a: effect of gender on exploration. b: effect of exploration on CPS.

Appendix D

Appendix D: Model 3 results and effect size κ² by country with migration background as predictor and number of interactions as mediator.

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ²
Australia	0.23	0.06	0.67	0.03	-0.09	0.06	0.15	0.04	0.07	0.06	0.07
Austria	-0.19	0.13	0.69	0.05	-0.27	0.13	-0.13	0.09	-0.41	0.13	0.02
Belgium	-0.34	0.11	0.77	0.03	-0.19	0.10	-0.26	0.09	-0.45	0.11	0.05
Brazil	-1.27	0.32	0.77	0.07	-0.22	0.76	-0.98	0.25	-1.20	0.75	0.10
Bulgaria	0.16	0.34	0.76	0.03	0.25	0.82	0.12	0.26	0.37	0.90	0.01
Canada	0.15	0.11	0.64	0.04	-0.34	0.08	0.09	0.07	-0.25	0.09	0.07
Chinese Taipei	0.70	0.20	0.65	0.06	-0.58	0.25	0.46	0.14	-0.12	0.30	NA
Colombia	-0.81	1.19	0.76	0.06	-0.70	2.46	-0.61	0.90	-1.31	1.76	0.03
Croatia	-0.10	0.10	0.81	0.04	-0.03	0.10	-0.08	0.08	-0.11	0.11	0.01
Czech Republic	-0.33	0.19	0.65	0.04	0.13	0.15	-0.21	0.13	-0.08	0.18	0.01
Denmark	-0.41	0.13	0.73	0.07	-0.16	0.12	-0.30	0.10	-0.46	0.13	0.05
Estonia	-0.01	0.15	0.76	0.06	-0.17	0.14	-0.01	0.11	-0.18	0.16	0.03
Finland	-0.53	0.18	0.72	0.04	0.12	0.13	-0.38	0.13	-0.25	0.10	0.00
France	-0.15	0.15	0.70	0.05	-0.26	0.14	-0.10	0.11	-0.37	0.16	0.03
Germany	-0.51	0.17	0.76	0.06	0.02	0.14	-0.39	0.13	-0.37	0.16	0.03
Hong Kong-China	0.10	0.10	0.78	0.05	-0.12	0.10	0.07	0.08	-0.04	0.11	0.00
Hungary	-0.04	0.30	0.71	0.07	-0.47	0.40	-0.03	0.21	-0.50	0.43	0.02
Ireland	0.06	0.19	0.72	0.06	-0.21	0.16	0.04	0.14	-0.17	0.17	0.04
Israel	0.03	0.12	0.82	0.04	0.20	0.09	0.03	0.10	0.22	0.13	0.03
Italy	-0.36	0.30	0.83	0.04	-0.23	0.30	-0.30	0.25	-0.53	0.31	0.01
Japan	0.75	0.52	0.68	0.04	-0.30	0.41	0.51	0.36	0.21	0.19	0.05
Macao-China	-0.08	0.08	0.67	0.08	-0.04	0.08	-0.05	0.06	-0.09	0.08	0.02
Malaysia	0.00	0.30	0.68	0.04	0.24	0.30	0.00	0.20	0.24	0.31	0.00
Montenegro	0.25	0.14	0.83	0.03	-0.10	0.10	0.21	0.12	0.12	0.14	0.01
Netherlands	-0.60	0.21	0.80	0.03	-0.27	0.10	-0.48	0.17	-0.74	0.20	0.01
Norway	-0.14	0.17	0.72	0.06	-0.46	0.18	-0.10	0.12	-0.56	0.16	0.03
Poland	-0.54	0.48	0.75	0.05	1.22	0.12	-0.40	0.36	0.82	0.40	NA
Portugal	0.08	0.16	0.81	0.05	-0.15	0.18	0.06	0.13	-0.09	0.18	0.04
Russian Federation	-0.18	0.17	0.67	0.05	-0.07	0.13	-0.12	0.11	-0.19	0.16	0.00
Serbia	-0.41	0.25	0.72	0.06	0.13	0.17	-0.30	0.17	-0.17	0.15	0.01
Shanghai-China	-1.38	0.72	0.58	0.06	-0.15	0.35	-0.80	0.42	-0.95	0.53	0.04
Singapore	0.06	0.11	0.59	0.06	0.04	0.11	0.04	0.06	0.08	0.11	0.03
Slovak Republic	-0.13	0.51	0.64	0.06	0.47	0.24	-0.08	0.33	0.38	0.30	0.02

Slovenia	-0.23	0.13	0.84	0.05	-0.04	0.19	-0.20	0.11	-0.23	0.18	0.10
Spain	-0.38	0.14	0.90	0.04	0.00	0.21	-0.34	0.13	-0.34	0.23	0.05
Sweden	-0.10	0.14	0.84	0.04	-0.17	0.13	-0.08	0.12	-0.26	0.15	0.03
Switzerland	-0.18	0.49	0.80	0.06	-0.32	0.25	-0.15	0.39	-0.46	0.34	0.02
Turkey	0.18	0.44	0.64	0.04	-0.14	0.28	0.11	0.28	-0.02	0.49	0.00
United Arab Emirates	0.49	0.07	0.74	0.04	0.07	0.07	0.36	0.05	0.43	0.07	0.15
United Kingdom	-0.14	0.13	0.62	0.07	-0.35	0.16	-0.09	0.08	-0.44	0.16	0.00
United States of America	0.34	0.13	0.59	0.07	-0.42	0.11	0.20	0.08	-0.22	0.10	0.05
Uruguay	0.13	0.36	0.79	0.03	0.25	0.43	0.10	0.29	0.35	0.35	0.02

Note. a: effect of migration on interaction. b: effect of interaction on CPS. NA indicates that κ^2 could not be calculated due to a zero covariance between migration background and interaction.

Appendix E

Appendix E: Model 4 results and effect size κ^2 by country with migration background as predictor and amount of exploration as mediator.

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ^2
Australia	0.34	0.09	0.43	0.06	-0.10	0.06	0.15	0.04	0.05	0.06	0.07
Austria	0.05	0.16	0.51	0.06	-0.34	0.13	0.03	0.08	-0.32	0.12	0.02
Belgium	-0.24	0.13	0.52	0.06	-0.22	0.12	-0.12	0.07	-0.34	0.11	0.05
Brazil	-2.79	0.50	0.33	0.12	-0.24	0.75	-0.92	0.40	-1.16	0.69	0.10
Bulgaria	-0.22	0.41	0.41	0.07	0.57	0.88	-0.09	0.17	0.48	0.88	0.01
Canada	0.38	0.09	0.36	0.09	-0.38	0.09	0.14	0.05	-0.25	0.08	0.07
Chinese Taipei	0.00	0.40	0.50	0.08	-0.12	0.32	0.00	0.20	-0.12	0.31	NA
Colombia	-1.19	0.61	0.70	0.10	-0.04	2.23	-0.83	0.46	-0.87	1.89	0.03
Croatia	-0.08	0.12	0.48	0.06	-0.02	0.10	-0.04	0.06	-0.06	0.10	0.01
Czech Republic	-0.18	0.20	0.39	0.05	0.04	0.15	-0.07	0.08	-0.03	0.16	0.01
Denmark	-0.25	0.24	0.58	0.25	-0.25	0.20	-0.14	0.19	-0.39	0.13	0.05
Estonia	0.20	0.17	0.52	0.08	-0.21	0.17	0.10	0.09	-0.11	0.15	0.03
Finland	0.03	0.14	0.28	0.08	-0.11	0.08	0.01	0.04	-0.10	0.09	0.00
France	0.18	0.16	0.41	0.08	-0.38	0.14	0.08	0.07	-0.30	0.14	0.03
Germany	-0.24	0.23	0.39	0.17	-0.16	0.18	-0.09	0.11	-0.26	0.15	0.03
Hong Kong-China	0.02	0.10	0.48	0.08	-0.06	0.10	0.01	0.05	-0.06	0.10	0.00
Hungary	-0.53	0.72	0.32	0.50	-0.20	0.47	-0.17	0.46	-0.38	0.40	0.02
Ireland	0.28	0.23	0.37	0.12	-0.25	0.17	0.10	0.09	-0.14	0.15	0.04
Israel	0.16	0.15	0.48	0.08	0.16	0.11	0.08	0.07	0.23	0.12	0.03
Italy	0.02	0.30	0.51	0.16	-0.52	0.26	0.01	0.15	-0.51	0.30	0.01
Japan	1.56	0.47	0.47	0.06	-0.47	0.24	0.73	0.24	0.26	0.24	0.05
Macao-China	-0.07	0.10	0.47	0.07	-0.04	0.08	-0.03	0.04	-0.07	0.08	0.02
Malaysia	-0.10	0.31	0.41	0.06	0.29	0.32	-0.04	0.13	0.25	0.29	0.00
Montenegro	0.05	0.22	0.46	0.07	0.08	0.14	0.02	0.10	0.10	0.12	0.01
Netherlands	0.07	0.21	0.27	0.10	-0.60	0.14	0.02	0.05	-0.58	0.16	0.01
Norway	0.23	0.20	0.34	0.11	-0.52	0.18	0.08	0.06	-0.44	0.16	0.03
Poland	0.05	1.43	0.46	0.12	0.82	0.57	0.02	0.66	0.84	0.30	NA
Portugal	0.23	0.21	0.57	0.08	-0.20	0.20	0.13	0.12	-0.07	0.18	0.04
Russian Federation	0.01	0.19	0.26	0.08	-0.16	0.14	0.00	0.05	-0.15	0.15	0.00
Serbia	-0.06	0.14	0.43	0.06	-0.10	0.15	-0.02	0.06	-0.13	0.14	0.01
Shanghai-China	-0.99	0.59	0.40	0.10	-0.51	0.35	-0.40	0.24	-0.91	0.46	0.04
Singapore	-0.15	0.14	0.39	0.13	0.15	0.12	-0.06	0.07	0.08	0.11	0.03
Slovak Republic	-0.65	0.23	0.33	0.08	0.49	0.34	-0.21	0.08	0.28	0.30	0.02
Slovenia	-0.32	0.21	0.79	0.12	0.00	0.24	-0.25	0.18	-0.25	0.18	0.10

Spain	-0.26	0.20	0.54	0.10	-0.10	0.21	-0.14	0.12	-0.24	0.23	0.05
Sweden	0.14	0.22	0.43	0.16	-0.26	0.15	0.06	0.11	-0.20	0.14	0.03
Switzerland	0.27	0.32	0.51	0.07	-0.51	0.32	0.14	0.16	-0.37	0.28	0.02
Turkey	0.08	0.39	0.34	0.07	-0.05	0.36	0.03	0.13	-0.02	0.45	0.00
United Arab Emirates	0.54	0.10	0.53	0.06	0.06	0.08	0.29	0.06	0.34	0.07	0.15
United Kingdom	-0.01	0.16	0.51	0.08	-0.44	0.17	0.00	0.08	-0.44	0.16	0.00
United States of America	0.36	0.14	0.31	0.08	-0.32	0.11	0.11	0.05	-0.21	0.10	0.05
Uruguay	0.66	0.40	0.31	0.07	0.16	0.32	0.20	0.13	0.36	0.33	0.02

Note. a: effect of migration on exploration. b: effect of exploration on CPS. NA indicates that κ^2 could not be calculated due to a zero covariance between migration background and exploration.