UNIVERSITÄT
DES
SAARLANDES

Universität des Saarlandes

Fakultät für Empirische Humanwissenschaften
und Wirtschaftswissenschaft
Fachrichtung Bildungswissenschaften

Predicting Educational Success - What's Beyond Intelligence

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften
der Fakultät HW
Bereich für Empirische Humanwissenschaften
der Universität des Saarlandes

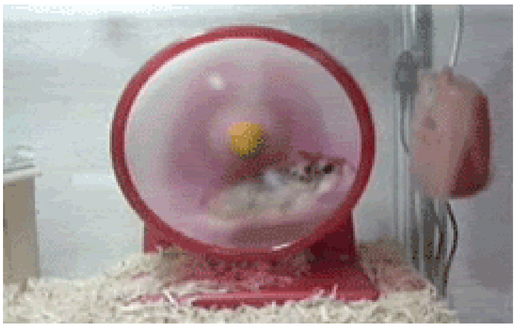
Vorgelegt von:
Dipl.-Psych. Christin Lotz
aus Ilmenau
Saarbrücken, 11. April 2018

Der Dekan:
Prof. Dr. Cornelius König

Berichterstatter:
Prof. Dr. Jörn R. Sparfeldt, Universität des Saarlandes
Prof. Dr. Joachim Wirth, Ruhr-Universität Bochum
Prof. Dr. Frank Spinath, Universität des Saarlandes

Tag der Disputation: 06.07.2018

Science:



Sometimes you spin the wheel,

sometimes it spins you.

... and sometimes you're just dizzy and not sure what happened.

Acknowledgement

First of all, I would like to express my gratitude to my Professor Jörn Sparfeldt for the aspiring guidance, useful comments and invaluable support throughout the whole process of my dissertation project. Back in school, my teachers repeatedly wrote on my report cards that I did not use my full potential. Possibly, the fist time in my life, this has changed during the last six years, as you pushed me to my limits – and beyond.

Also, I would like to thank my parents who have taught me to value education. Although you did not fully understood my works, you always encouraged me that this lousy-paid job in educational science will result in something useful that might be important for improving society.

Furthermore, I would like to thank Rebecca Schneider. You were the best coworker I ever had and I could wish for. Throughout our 4 years of shared success and despair, laughter and agony, disappointment and delight, the bad was only half as bad but the good was twice as good, as we had it in common. In you, I found one of my best friends during my time in Saarbrücken and an excellent travel companion. Thank you for the loads of fun we had on our numerous conference travels within Germany and, especially, abroad. I will never forget that it was you who taught me how to order grilled doggies in Russian. Saved my life!

Many thanks also go to my former and current colleagues. Having such fantastic coworkers, I noticed how important it is to work in an companionable team that provided me with helpful discussions and constructive criticism. Especially, I would like to mention Johannes Schult. Without you, I would have never mastered M*plus*. Thus, your patience in teaching statistics and your bright ideas in case of non-positive definite co-variance matrices extraordinarily contributed to this dissertation. Moreover, the ”ANOVA mit gerichteten Kontrasten” Song really persisted in my head.

In addition, I would like to thank my co-authors for all the helpful remarks, advises, and discussions. My particular gratitude is expressed to Ronny Scherer who supervised me during my research visit at the university of Oslo.

Relating thereto, I express my gratitude to CEMO for the interesting talks I had with the CEMO members and for the excellent infrastructure I was allowed to use during my research visit. At CEMO, I received a vivid impression about how research is conducted if a university is not suffering from extensive budged cuts.

Moreover, special thanks go to Johannes Hellenbrand, Sabrina Navratil, Tim Kühl, Tobias Ringeisen, and Anja Meißner for making most of the conferences feel like a huge class trip. With your company, exhausting days became fun (and sometimes educational sightseeing experiences), whereas gala dinners became unforgettable night outs.

Last but not least, I want to express my deepest gratitude to Sascha Ludwig. As you correctly mentioned once, it must be real love if someone keeps listening with interest to work-related problems over years! You were the one who supported me most constantly throughout this entire dissertation process. You gave me valuable pieces of advice from your objective but involved and educated point of view. Moreover, it was your magic computer-science-power that extracted the CPS log-file data. Thus, without your expertise and involvement, study 2 would have hardly ever made it to Intelligence. But most importantly, thank you for sending me cute cat videos or science-related memes when I needed distraction or cheering ups. YOU were my work-life-balance! <3 <3 <3

Preface

Theoretical considerations concerning the design of the data set were made by Jörn Sparfeldt and Samuel Greiff. Thus, the data set that built the basis of this dissertation project is owned by Jörn Sparfeldt and Samuel Greiff. It was their main objective to answer the research questions of study 1 with this data set. At the point of assembling the specific test battery, I got involved in the project and subsequently co-organized the data collection.

With reference to the three empirical studies, the shares were as follows: Concerning study 1, the main research questions were developed by Jörn Sparfeldt and Samuel Greiff. I conducted the analyses and wrote large parts of the manuscript with contributions of the two co-authors.

With reference to study 2, the idea to extract CPS log-files and to analyze them by means of latent growth curve models was generated by myself. I conducted the analyses during a research visit at the Center for Educational Measurement at the University of Oslo under the supervision of Ronny Scherer and Jörn Sparfeldt. Large parts of the corresponding manuscript were written by myself with contributions of the three co-authors.

Concerning study 3, it was Jörn Sparfeldt's intention to examine an expectancy-value-related research question and, therefore, motivational measures were included in the data set. The idea to specifically analyze the differential relevance of motivation and intelligence by means of reparameterization was generated by myself. Accordingly, I conducted the analyses and wrote large parts of the manuscript with contributions of the two co-authors.

I thank Jörn Sparfeldt and Samuel Greiff for giving me the opportunity to use this excellent data set for conducting the three empirical studies of my dissertation project.

Contents

List of Tables

List of Figures

List of Abbreviations

BIS	Berlin Intelligence Structur Model
CFI	Comparative Fit Index
CFT	Culture Fair Test
CPS	Complex Problem Solving
df	Degrees of Freedom
DISC	Differential Self-concept
GPA	Grade Point Average
HOTAT	Hold one Thing at a Time
LGCM	Latent Growth Curve Model
M	Mean
MCS	Multiple Complex Systems
MI	Modification Indices
NOTAT	Vary no Thing at a Time
OECD	Organization for Economic Co-operation and Development
PISA	Program for International Student Assessment
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
SEM	Structural Equation Modeling
TLI	Tucker-Lewis-Index
VOTAT	Vary one Thing at a Time

List of Publications

This dissertation project is based on three scientific articles that are published in peer-reviewed journals. Articles are available online through the respective publishing company.

1. Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex Problem Solving in Educational Contexts — Still something beyond a "good *g*"? *Intelligence, 59,* 127–138.
2. Lotz, C., Scherer R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence, 64,* 98–112.
3. Lotz, C., Schneider, R., & Sparfeldt, J. R. (2018). Are intelligence and motivation differentially relevant for scholastic competence tests and grades in mathematics? *Learning and Individual Difference, 65*, 30-40.

In the following, I will use the term *we* when referring to these studies that my co-authors and I have conducted together.

Summary

In educational contexts, intelligence is regarded as an extraordinary powerful psychological predictor of scholastic achievement. However, substantial amounts of variance are left unexplained, if only intelligence is considered as a predictor. Therefore, we investigated the incremental validity of selected cognitive, strategic-behavioral, and motivational constructs beyond broadly operationalized intelligence for educational success assessed by scholastic competence tests and grades in the key school subjects mathematics and German. To achieve this, three separate studies were conducted.

In study 1, the incremental validity[1] of complex problem solving (CPS) was examined. CPS claims to tap unique cognitive aspects that are not assessed by conventional intelligence tests. Accordingly, prior studies revealed promising results for the prediction of scholastic achievement. However, because most of these studies only assessed narrow measures of intelligence (mostly only figural reasoning), a more thorough investigation of the CPS increment beyond comprehensive measures of intelligence was needed. Results of study 1 indicated a huge overlap between broadly assessed intelligence and CPS. Moreover, intelligence prevailed as the superior predictor which was evidenced by significantly higher path coefficients and higher increments on scholastic competence tests and grades in mathematics as well as in German. Contrarily, CPS only revealed small unique effects on the mathematics competence test.

As CPS tests provide computer-generated log-files about the students' behaviors while solving complex problems, we analyzed the interactive exploration behavior during the CPS knowledge acquisition phase in study 2. Accordingly, we investigated how the optimal strategic exploration behaviors VOTAT (vary one thing at a time) and NOTAT (vary no thing at a time) were applied and adapted across a CPS task set. Moreover, we examined whether intelligence facilitates more effective strategy use. Results of discontinuous latent growth curve models (LGCM) showed that students applied the effective strategic behaviors, i.e. VOTAT and NOTAT, and that they were able to flexibly adapt their

[1] In this dissertation, the term incremental validity is used in the sense of cross-sectional statistical prediction; i.e., explaining variance in a criterion.

strategy use to occurring changes in the task type. Moreover, intelligence manifested itself in a more effective strategy use as more intelligent students applied VOTAT and NOTAT with higher frequencies and adapted them with a steeper gradient if these behaviors were effective, but in turn, they applied them less often if these behaviors were not effective. Supplementary analyses examined the incremental validity of the strategic behaviors for educational success beyond intelligence. Whereas VOTAT exhibited no substantial unique effects, NOTAT revealed small but substantial increments on the scholastic competence test and grades in mathematics.

In Study 3, the focus was shifted towards motivational variables. Ample evidence indicated that intelligence, academic self-concept, and academic interest exhibit a differential relation pattern, depending on which achievement indicator was used as the criterion of educational success. Whereas intelligence seems to be more important for predicting scholastic competence tests, self-concept and interest seem to be of higher importance for predicting grades. However, this differential prediction pattern was only numerically described in prior studies and awaited its examination by advanced statistical methods. Accordingly, we applied reparamerization in combination with inferential-statistical tests to compare the standardized path coefficients. As expected, intelligence revealed to be the superior predictor for scholastic competence tests, whereas self-concept was superior for predicting grades. Moreover, self-concept showed considerable incremental effects beyond intelligence on grades. Nevertheless, the incremental validity of interest seemed to be based on the huge overlap with self-concept because as we controlled for their common variance, interest was non-predictive for the criteria of educational success.

To conclude, the results of this dissertation project confirmed intelligence as one of the most prevalent psychological predictors of educational success. Whereas other cognitive or strategic-behavioral predictors seemed to be of rather negligible importance beyond comprehensive measures of intelligence, especially self-concept revealed to be of extraordinary relevance for predicting grades.

Zusammenfassung

Die Vorhersage von Schulerfolg ist seit jeher einer der wichtigsten Forschungsbereiche der Pädagogischen Psychologie. Die allgemeine Intelligenz nimmt hierbei eine sehr bedeutsame Rolle ein. Allerdings reicht die allgemeine Intelligenz, trotz ihrer starken Vorhersagekraft, nicht als einziger Prädiktor von schulischen Leistungen aus, da große Anteile der Schulerfolgsvarianz von ihr nicht aufklärt werden. Die Betrachtung weiterer Einflussfaktoren erscheint daher äußerst lohnenswert. Anhand von drei separaten Studien wurde in der vorliegenden Dissertation die inkrementelle Validität ausgewählter kognitiver, strategisch-behavioraler und motivationaler Konstrukte bei der Vorhersage von Schulerfolg, erfasst durch standardisierte Schulleistungstests und Noten in den Kernfächern Mathematik und Deutsch, untersucht.

In Studie 1 wurde die inkrementelle Validität des komplexen Problemlösens (KPL) erforscht. Von KPL-Tests wird angenommen, dass sie bestimmte kognitive Facetten erfassen, die von herkömmlichen Intelligenztests nicht abgedeckt werden. Bisherige Studien, die auf eine inkrementelle Validität des komplexen Problemlösens über Intelligenz hinaus schließen ließen, erfassten allerdings meist nur figurales reasoning als eine Facette der Intelligenz, nicht aber die allgemeine Intelligenz im Sinne von *g*. Die Ergebnisse von Studie 1 zeigten eine sehr hohe Korrelation zwischen der breiten Intelligenzoperationalisierung und KPL. Weiterhin wurde bestätigt, dass die allgemeine Intelligenz im Vergleich zum komplexen Problemlösen der statistisch bedeutsamere Prädiktor von allen untersuchten Schulerfolgsmaßen war. Außerdem wies die Intelligenz bedeutsame Inkremente für alle Schulerfolgsmaße über KPL hinaus auf. Im Gegensatz dazu ergab das komplexe Problemlösen nur ein vergleichsweise kleines Inkrement über die allgemeine Intelligenz hinaus für den Schulleistungstest in Mathematik.

KPL-Testinstrumente zeichnen dank ihrer computerisierten Erfassung Logfiles über die Verhaltensweisen bei Lösen komplexer Probleme auf. In Studie 2 nutzten wir dieses Potential, um das strategische Explorationsverhalten während der interaktiven Wissenserwerbsphase in KPL-Aufgaben zu erforschen. Wir untersuchten (a) die Anwendungshäufigkeit der optimalen strategischen Explorationsverhaltensweisen VOTAT und NOTAT, (b) deren flexible Anpassung an

sich verändernde Aufgabenanforderungen über eine Aufgabenserie hinweg und (c) inwieweit höhere Intelligenz mit effektiverem Strategieeinsatz einherging. Latente Wachstumskurvenmodelle ergaben, dass die Schüler effektive Strategien anwandten und diese auch flexibel an wechselnde Aufgabenanforderungen anpassen konnten. Darüber hinaus zeigte sich, dass intelligentere Schüler VOTAT und NOTAT häufiger anwandten und schneller anpassten, wenn diese strategischen Verhaltensweisen effektiv waren. Waren die Verhaltensweisen nicht effektiv, setzten sie die Schüler seltener ein. Intelligenz schien sich also im Problemlöseprozess zu manifestieren, indem sie mit dem optimalen Einsatz von effektiven Verhaltensweisen einherging. Zusatzanalysen, welche die inkrementelle Validität der strategischen Verhaltensweisen für Schulerfolg untersuchten, ergaben für VOTAT keine nennenswerte zusätzliche Varianzaufklärung über Intelligenz hinaus. Allerdings zeigten sich geringe inkrementelle Effekte für NOTAT bei der Vorhersage des standardisierten Schulleistungstests und der Noten in Mathematik über Intelligenz hinaus.

In Studie 3 fokusierten wir die differentielle Bedeutsamkeit von Intelligenz und den motivationalen Variablen Selbstkonzept und Interesse. In der bisherigen Literatur zeigten sich Hinweise darauf, dass Intelligenz von höherer Bedeutung bei der Vorhersage von Schulleistungstests sei, während das akademische Selbstkonzept und das akademische Interesse bei der Vorhersage von Schulnoten von höherer Bedeutung zu sein scheint. Diese differentielle Relevanz wurde bislang allerding nicht mit state-of-the-art Methoden, die explizit für den inferenz-statistischen Vergleich standardisierter Pfadkoeffizienten entwickelt wurden, auf ihre statistische Bedeutsamkeit überprüft. Die Ergebnisse von Studie 3 zeigten, dass bei der Vorhersage von Schulleistungstests die Intelligenz der wichtigere Prädiktor war, während bei der Vorhersage von Schulnoten das akademische Selbstkonzept der bedeutendere Prädiktor war. Weiterhin konnte das Selbstkonzept beträchtliche Varianzanteile in den Schulnoten über Intelligenz hinaus aufklären. Die prädiktive Validität des akademischen Interesses scheint allerdings vollkommen auf die geteilte Varianz mit dem Selbstkonzept zurückführbar zu sein, da das Interesse keine bedeutsamen Zusammenhänge zu Schulleistung mehr aufwies, sobald für diese geteilte Varianz kontrolliert wurde.

Zusammenfassend kann festgehalten werden, dass der herausragende Stellenwert der Intelligenz bei der Vorhersage von Schulerfolg in diesem Dissertationsprojekt aufs Neue untermauert wurde. Während die Bedeutung anderer kognitiver und strategisch-behavioraler Konstrukte über die allgemeine Intelligenz hinaus eher vernachlässigbar erscheint, erwies sich im Speziellen das akademische Selbstkonzept als überaus wichtiger Einflussfaktor für Schulnoten.

Chapter 1

Introduction

Investigating the determinants of educational success is one of the oldest and at the same time one of the hardest challenges for educational psychology because high scholastic achievements are regarded as desirable educational outcomes. Typically, scholastic competence tests or teacher given grades are used to assess scholastic achievement as they signify to which extent specific educational goals were accomplished (Steinmayr, Meißner, Weidinger, & Wirthwein, 2014; Steinmayr, Sauer, & Gamsjäger, 2018). Although both measures of educational success tend to correlate substantially, they seem to form distinct measures (e.g., r = .40 Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005.

With reference to grades, they typically assess the students' achievement in a specific school subject, but they are also influenced by non-ability aspects such as motivation, personality, and working behavior of the students, as well as the teachers' perception of these aspects. Thus, teacher given grades are a rather heterogeneous measures of educational success (Harlen, 2005; Willingham, Pollack, & Lewis, 2002). In connection therewith, their psychometric properties concerning objectivity, reliability, and validity are not undisputed and several empirical studies reported vulnerabilities in the assignment of grades (see Birkel & Tarnai, 2018). However, although the psychometric properties of teacher given grades are not perfect, they contain manifold important functions in the German educational system (Birkel & Tarnai, 2018; Füssel & Leschinsky, 2008). Their pedagogical functions are (a) socialization functions by familiarizing students with norms, (b) report functions by giving students evaluative feedback, by informing parents about the scholastic achievement of their children, and by providing information on eventually required educational support, and (c) incentive functions as good grades are supposed to motivate, whereas bad grades are supposed to discipline students. Moreover, grades have societal functions like (a) authorization functions by legitimizing the educational actions of the teachers, (b) control function by making the effects of political or educational interventions transparent, and (c) selection or allocation function by allowing the promotion to the next school year or the assignment of an university place on the basis of actual achievements and not on the basis of, for example, social background.

On the contrary to grades, scholastic competence tests play a minor role in the German educational system (Füssel & Leschinsky, 2008). Typically, standardized tests are not generated by the students' corresponding teachers, but they are compiled and evaluated by specific institutions such as the OECD (Organization for Economic Co-operation and Development) or by researchers. Thus, scholastic competence test posit some advantages compared to teacher given grades (Lissmann, 2018). For example, they are carefully constructed and their psychometric properties meet common standards. Moreover, they are standardized so that the individual test score can be compared to curricular benchmarks or a particular group of reference. Thus, scholastic competence tests are often employed to investigate learning abilities or learning results. However, although scholastic competence tests comprise satisfying psychometric properties and many important functions, they do not serve as a direct source of feedback for the students, as the tests are not evaluated by the students' corresponding teachers and are not graded at all. These aspects also contribute to the low importance of scholastic competence tests in Germany.

When the question arises which variables determine educational success, intelligence is widely regarded as one of the most prevalent psychological predictors (e.g., Brühwiler & Helmke, 2018; Neisser et al., 1996). Nevertheless, a high level of intelligence is "only a necessary but never a sufficient condition" for doing well in educational contexts (Jensen, 1998b, p. 122). Thus, focusing on only intelligence to explain success or failure in educational settings might lead to misspecifications or too short-sighted conclusions. Because of the multiple determination of educational success, a variety of variables might contribute substantially to the prediction of educational success beyond intelligence. Moreover, these variables interact with intelligence and among each other in a complex framework. Accordingly, examining the incremental validity of other predictors beyond intelligence seems rewarding for several reasons: First, conventional intelligence tests were occasionally criticized due to their supposed mismatch to the demands and requirements of everyday life. The challenges of real-life seem to be much more complex than what could be assessed by conventional intelligence tests. Therefore, the construct of complex problem solving might have the potential to complement or replace conventional intelligence tests and to explain educational success even above and beyond intelligence. Second, conventional intelligence tests typically merely provide final performance scores but no process measures. CPS tests provide such process measures that can be used for a detailed analysis of strategic behaviors during the problem solving process. This might open up new possibilities to investigate the incremental validity of strategic-behavioral variables for educational success beyond intelligence. Third, students frequently do not perform as good as they are actually capable of. Considering motivational variables such as self-concepts or interests might be key elements for predicting educational success beyond intelligence.

Thus, in this dissertation project, the incremental validity of cognitive, strategic-behavioral, and motivational variables for predicting educational success beyond intelligence is highlighted. First, the performance measures of CPS as a cognitive variable are regarded. Second, the strategic-behavioral process measures of CPS that tap effective strategic behaviors for exploring unknown problem spaces are considered. Third, the focus is shifted towards the motivational variables academic self-concept and interest.

Accordingly, in the first study of this dissertation project the incremental validity of CPS performance measures for educational success beyond intelligence was examined. Although predicting educational success is the genuine purpose of intelligence tests and most conventional intelligence tests fulfill this purpose considerably well (Neisser et al., 1996), CPS is discussed as a viable alternative or complement to conventional intelligence tests as CPS tests claim to tap aspects that are only partially covered by conventional intelligence tests (Greiff, Wüstenberg, & Funke, 2012; Wüstenberg, Greiff, & Funke, 2012). These aspects, comprising cross-curricular skills such as planning and implementing actions, model building, and self-regulation seem to become of significant importance in today's educational system and society (Funke, 2010) as they are ranged among the 21st century skills (Pellegrino & Hilton, 2012). Thus, a more thorough examination of the incremental validity of CPS for educational success beyond intelligence seems fruitful because the educational system appears to be in a transitory state in which these so-called 21st century skills are ascribed increasing importance. Moreover, CPS tests revealed promising first results in prior studies, in which CPS incrementally predicted scholastic achievement beyond tests of figural reasoning as the most prototypical facet of intelligence (e.g., Greiff et al., 2012; Wüstenberg et al., 2012). Nevertheless, the examination of the incremental effects of CPS beyond comprehensive measures of intelligence for scholastic competence tests and grades is still pending.

In the second study of this dissertation project, a more detailed examination of the CPS process measures was targeted. Contemporary CPS tests provide process measures of every single action during the problem solving process due to their computer-based assessment. In contrast to conventional intelligence tests, they feature a phase in which the problem space could be freely and interactively explored by the problem solver. By analyzing the computer-generated log-files, effective strategic exploration behaviors such as VOTAT (Tschirgi, 1980) or NOTAT can be examined and their application across a task set can be investigated. Moreover, examining the relations between effective strategic behaviors and intelligence as well as their incremental validity for educational success beyond intelligence might reveal completely new insights into the utility of strategic-behavioral CPS process measures.

In the third study of this dissertation project, the focus was shifted towards motivational variables because in prior research, there was ample evidence that motivation also represents an important determinant of educational success (e.g., Hattie, 2009; Valentine, DuBois, & Cooper, 2004). Especially regarding the key concepts of expectancy-value-theory, self-concept and interest, there are hints for a differential relevance of cognitive and motivational variables, depending on which indicator of educational success was used. Whereas intelligence seems to be of higher importance for scholastic competence tests, motivational variables seem to be more important for teacher-given grades (e.g., Helmke, 1992; Jansen, Lüdtke, & Schroeders, 2016; Steinmayr & Meißner, 2013). However, so far this differential relation pattern was only numerically described, but it was not inferential-statistically tested by advanced and state-of-the-art methods. Thus, a statistical examination of the differential prediction pattern might highlight the significance of motivational variables, especially for grades. Thereby, the importance of distinguishing between scholastic competence tests and grades as indicators of scholastic achievement is emphasized.

To conclude, in this dissertation project, we pursued the aim of examining the incremental validity of cognitive, strategic-behavioral, and motivational variables beyond intelligence on educational success assessed via scholastic competence tests and grades in the two key school subjects mathematics and German as the native language.

Chapter 2

Theoretical Framework

This chapter provides a brief summary about the theoretical framework that this dissertation project is based on. First, a general theoretical background concerning intelligence and its significance for educational success is provided. Second, the cognitive construct complex problem solving and its relations to intelligence and educational success is portrayed. Thereby, performance measures and process measures of complex problem solving are distinguished. Third, the two key concepts of motivation, i.e. self-concept and interest, are introduced and their incremental validity for educational success is considered.

2.1 Intelligence and Educational Success

Intelligence is one of the most prevalent psychological constructs and of extraordinary importance for predicting educational success. Therefore, the following section addresses the correlational structure of intelligence and describes important guidelines for assessing a broad general factor of intelligence. Furthermore, the predictive validity of intelligence for educational success is detailed, thereby differential effects for grades and scholastic competence tests as well as for the mathematics and the verbal domain are highlighted.

2.1.1 Intelligence

In research on intelligence, it is a well-known and often replicated fact that all cognitive achievement subtests tend to be positively correlated, even if they measure seemingly different cognitive abilities (e.g., verbal or figural skills). This correlational pattern is called positive manifold (Spearman, 1904) and implies the existence of a general factor of mental abilities, symbolized as *g*, which was confirmed in hundreds of factor-analytic studies (Carroll, 1993). Besides this positive manifold, there are usually specific patterns of intercorrelations among groups of tests that represent more specific group factors (e.g., Carroll, 2003; McGrew, 2009). Today, a *g*-based factor hierarchy with a general factor of intelligence (*g*) at the apex and various more specific ability (group) factors arrayed below it, is the most accepted view of the correlational structure of intelligence (Neisser et al., 1996).

Although the *g*-factor is frequently used in contemporary research on intelligence, not any set of intelligence subtests that contains some *g*-variance necessarily represents what Jensen and Weng (1994) called a ”goog *g*”. Because the strength of *g* is significantly affected by the composition of the test battery, the authors proposed that the quality of an extracted *g*-factor depends upon (a) the number of subtests in the test battery, (b) the reliability of the subtests, (c) the diversity of the subtests, and (d) the comparable and balanced representation of the diverse subtests within the test battery.

Regarding (a), the number of subtests to be included in the test battery, a purely psychometric perspective would suggest 'the more subtests the better' because an increasing number of (reliable and valid) subtests would increase the quality of *g* by reducing the amount of psychometric sampling error. From an empirical point of view, a test battery should contain at least ten subtests in the composite to lead to stable estimates of *g* in factor analyses (Jensen & Weng, 1994; Ree & Earles, 1991). This suggestion was widely taken into account in current intelligence research, as a review with 144 unique analyses of intelligence test batteries showed: on average, 14 subtests with an interquartile range from 8 to 16 subtests were included in the test batteries (Reeve & Blacksmith, 2009, after excluding studies with three or less subtests).

With reference to (b), the reliability of the subtests, high reliability coefficients are without doubt of extraordinary importance when measuring any kind of construct, including intelligence. High measurement errors endanger the quality of the measurement and, therefore, the prerequisites for acceptable validity coefficients are not given (Wittmann, 1988).

Next to the number and the reliability of the indicators, (c) the extent of the diversity of the intelligence subsets is of high importance for assessing a comprehensive general factor. Empirical evidence indicated that specific ability factors typically only predict narrow content specific components of the criterion variance (Reeve, 2004). However, when predicting a broader criterion like educational success — instead of only specific abilities — a broader predictor is needed (Wittmann, 1988). Thus, a wide range of diverse intelligence subtests, representing different group factors, instead of highly similar indicators is recommended to receive an appropriate measure of *g*.

To determine the broadness of the intelligence test battery, the Berlin Model of Intelligence Structure (BIS; Jäger, Süß, & Beauducel, 1997, see Fig. 2.1) seems especially well-suited. In accordance with the mostly accepted intelligence structure model, the BIS model postulates a g-factor at the apex and several ability components below it. The ability components belong to two facets, resulting in a bi-facet structure with three content facets (verbal, numerical, and figural) and four operation facets (reasoning, memory, speed, and creativity). The cross-classifications of the operation and content facets yields in total 12 ability cells on a third level. Accordingly, this model can serve as a classification scheme for different intelligence tasks. For instance, a figural analogy task can be fitted in the BIS cell which results from the crossing between the figural content facet and the reasoning operation facet. Thus, this classification scheme helps to determine whether an intelligence test battery is diverse: The more crossings are covered, the more diverse is the test battery.

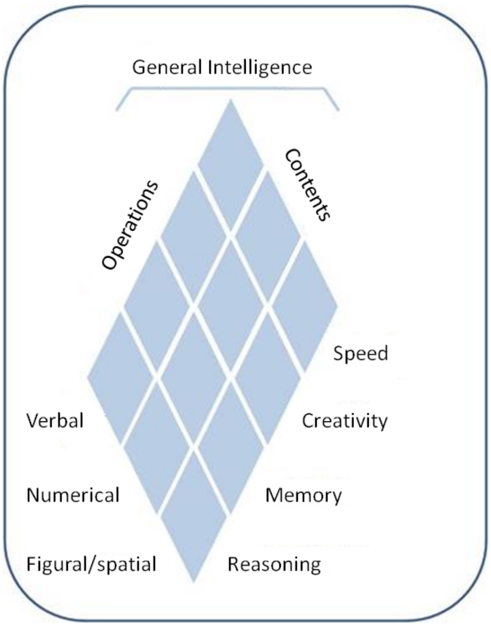


Figure 2.1: Berlin Intelligence Structure Model (BIS-model; according to Jäger et al., 1997). The four operation facets speed, creativity, memory, and reasoning as well as the three content facets verbal, numerical, and figural yield 12 ability cells. Their integral yields general intelligence.

The last suggestion of Jensen and Weng (1994) focused on (d), the balanced representation of the diverse subtests in the test battery. Again, the BIS model serves as a classification scheme to detect possible overweights or imbalances in intelligence test batteries. For example, assessing mostly subtests of the figural content facet causes a figural overweight in the estimation of g. Analogously, a substantial imbalance in one of the operation facets causes, for example, that reasoning becomes too dominant in *g*. As a consequence, the test battery would rather only measure reasoning instead of general intelligence (Stankov, 2012).

Thus, according to Jensen and Weng (1994), it can be concluded that the quality of an extracted *g*-factor strongly depends on the composition of the test battery. At first glance, this conclusion seems to contradict the principle of the indifference of the indicator (Spearman, 1923). According to this theorem, the precise content of intelligence test batteries is unimportant for the purposes of identifying *g*. Because of the positive manifold, all indicators load on the general factor to some extent and in the composite, they universally accumulate into an estimate of *g*. However, this independence from the specific operationalization is only a relative independence because the indifference of the indicator is only given as long as the intelligence test battery is large and diverse (e.g., Brody, 2000; Rost, 2013). If, for example, only measures of figural reasoning are assessed, the relations to *g* are still high but, nevertheless, the estimate is not necessarily representing the general factor of intelligence (Johnson, Nijenhuis, & Bouchard, 2008). Therefore, assessing a comprehensive measure of intelligence is an important prerequisite for robust examinations of the incremental validity of intelligence for educational success.

2.1.2 Intelligence as a Predictor of Educational Success

There is ample evidence that intelligence, in the sense of *g*, is one of the best psychological predictors for educational success and accounts for more variance than any other single psychological factor independent of *g* (e.g., Jensen, 1998b; Kuncel, Hezlett, & Ones, 2004; Mackintosh, 2011). The close relation between cognitive abilities and educational outcomes is based on the *g*-demandingness of learning and scholastic achievement itself. For example, students have to grasp concepts and meanings, learn to deal with novel material, make distinctions, recognize patterns, or transfer previously learned knowledge and skills to new situations – which is all intrinsic to *g* (Jensen, 1998b).

Educational success is typically assessed by teachers' grades or scholastic competence tests (Steinmayr et al., 2018). Interestingly, intelligence seems to display a differential relevance for these two achievement indicators, as the relations with standardized test scores seem to be higher than the relations with grades. Referring to grades, meta-analytic results revealed that the average validity coefficient between intelligence and grades is $\rho = .54$ (Roth et al., 2015) and further confirmed the close relation of about .50 that was proposed by Jensen (1998b). Correlations of this magnitude were also repeatedly reported for German samples (e.g., Brühwiler & Helmke, 2018; Helmke & Weinert, 1997; Kühn, 1987; Klauer & Sparfeldt, 2018; Mehlhorn & Mehlhorn, 1981). When considering composite measures of grades, the proportion of explained variance is even higher. For instance, a longitudinal study with a large and representative sample reported a substantial correlation of $r = .81$ between the g-factor of intelligence measured at age 11 and a general factor of scholastic achievement, consisting of final exam grades in six different school subjects at age 16 (Deary, Strand, Smith, & Fernandes, 2007). Comparable results were reported for similar large and representative international samples (Calvin, Fernandes, Smith, Visscher, & Deary, 2010: $r = .83$; Kaufman, Reynolds, Liu, Kaufman, & McGrew, 2012: $.81 < r < .86$). For a German sample, using the nested-factor approach, a general factor of intelligence correlated to $r = .44$ with a general factor of scholastic achievement, consisting of grades in six different school subjects (Valerius & Sparfeldt, 2015). In addition, also the nested specific verbal and specific numerical intelligence factors correlated substantially with the nested specific verbal and specific mathematics/scientific scholastic achievement factors ($r_{\text{verbal}} = .49$; $r_{\text{numerical}} = .51$).

Referring to scholastic competence tests, the correlation coefficients with intelligence tend to be numerically higher than the average intelligence-grade coefficients. Whereas, as mentioned, grades correlate with cognitive abilities to around .50 on average (Jensen, 1998b; Roth et al., 2015), the correlations between intelligence tests and the performance on standardized achievement tests usually range from about .60 for achievement subtest scores to .70 for composite scores on a manifest level (Naglieri & Bornstein, 2003). When considering latent coefficients, the relations were even closer (Baumert, Lüdtke, Trautwein, & Brunner, 2009; Frey & Detterman, 2004). One possible explanation for their very strong association might be the moderating role of intelligence during the cumulative process of knowledge acquisition whose results are measured by scholastic competence tests (Baumert et al., 2009). Moreover, a recent large-scale study provided empirical evidence for the described differential relevance of cognitive abilities for scholastic competence tests and grades across five school subjects in one sample (Jansen et al., 2016). For instance, a measure of figural reasoning revealed numerically higher regression coefficients on a mathematics competence test ($\beta = .56$) than on the mathematics grade ($\beta = .26$). Similar patterns were reported for the school subjects German, biology, chemistry, and physics.

When considering different domains of educational success, different school subjects pose different demands upon the students. Correspondingly, differential domain-specific relations to intelligence were reported in international (e.g., Calvin et al., 2010; Deary et al., 2007) and national samples (e.g., Brunner, 2008; Kühn, 1987). Apparently, dealing with geometrical figures in mathematics is more g-demanding than spelling in the native-language lessons or drawing a picture. Empirical evidence from large-scale studies fosters this assumption: Deary et al. (2007) reported correlations between intelligence and different school subjects that varied between $r = .77$ for mathematics, $r = .67$ for the native language lesson down to $r = .43$ for art and design. Thus, there seems to be a differential correlation pattern for different school subjects, whereas the correlation between *g* and mathematics was always the highest among all school subjects because mathematics (and science-related school subjects) draw to a stronger extend upon the central cognitive resources than other school subjects (Jensen, 1998b).

Taken together, intelligence has enormous conspicuous predictive validity for educational success. Concerning the indicators of educational success, scholastic competence test scores tend to correlated closer with intelligence than grades. Regarding the differential *g*-demands of various subjects, mathematics was found to have the highest *g*-loadings.

However, besides *g*'s power to explain large amounts of variance in scholastic achievement and to provide optimal prerequisites for educational success, a high level of intelligence is, nevertheless, "only a necessary but never a sufficient condition" for doing well in educational contexts (Jensen, 1998a, p.122). Thus, despite the excellent predictive validity of intelligence for educational success, considerable proportions of variance remain unexplained (Neisser et al., 1996), which implies that other variables might play an important role for predicting educational success beyond intelligence.

2.2 Complex Problem Solving and Educational Success

Complex problem solving is discussed as one of the most promising cognitive constructs to fill in the gap that is left next to intelligence when predicting educational success. Thus, the following section focuses on this construct by providing a short overview about the construct itself and its measurement approach. Moreover, the findings about the relations between CPS and intelligence as well as the incremental validity of CPS for educational success above and beyond intelligence are reviewed. Thereby, CPS performance measures and CPS process measures are differentiated.

2.2.1 Complex Problem Solving

Out of the attempts to conceptualize CPS as a psychological construct, Buchner (in Frensch & Funke, 1995) provided one of the most precise definitions:

> The successful interaction with task environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process. (p.14)

Moreover, five key criteria to determine complex problems were established by Dörner (1980a) and Putz-Osterloh (1981): (1) *complexity*: the problem consists of several variables, (2) *connectedness*: the variables need to be linked via numerous connections, (3) *dynamic*: the problem situation is changing itself over time, (4) *non-transparency*: not all necessary information to solve the problem are initially available, and (5) *polytelic*: solving the problem requires the optimization of more than one criterion.

Ever since the construct of CPS was established, it emerged as a vivid research topic. Accordingly, a vast variety of different CPS tests, dealing with topics like biology, medicine, economy, or ecology were developed over the last 40 years (for an overview see Fischer, Greiff, & Funke, 2017; Funke, 2006). Some of the CPS tests feature semantically rich environments. Of those, the most significant and frequently used tests are, for example, Lohausen (governing a little town; Dörner, Kreuzig, Reither, & Strohschneider, 1983), Moro (providing developmental aid to a small African semi-nomadic tribe; Dörner, Stäudel, & Strohschneider, 1986), Fire Fighting (protecting a city from approaching fires; Brehmer, 1987), or Tailorshop (managing a tailor shop; Dörner, 1979). Other CPS tests are based on more abstract environments as, for example, MultiFlux (Kröner, 2001) or MicroDYN (Greiff et al., 2012).

Especially the early-developed, semantically-rich, and realistic CPS tests demonstrated high face validity as complex real-life scenarios were simulated. However, precisely the semantic richness caused misconceptions on the side of the problem solvers. For instance, in the CPS test Tailorshop, it would be reasonable to assume a dependency between supply and demand of the raw material, which was, in fact, not implemented in the test. Accordingly, such non-correspondences between real-life regularities and CPS tests might cause problem solvers to fail in CPS tests in spite of proper representations of the problem space (Funke, 2006). One approach to avoid the uncontrolled influence of prior knowledge was to minimize the semantic content of the CPS tests. Those knowledge-lean and rather simple scenarios are called minimal complex systems because of their lower number of variables and relations between the variables (Greiff & Funke, 2009). In those scenarios, the rather few variables are labeled without deep semantic meaning. For example, in the MicroDYN item "handball" even an experienced handball trainer could not rely on prior knowledge or heuristics to determine how the three possible training types A, B, or C affect three different team characteristics (see below for a more detailed description of MicroDYN).

Another approach was to include a standardized learning phase into the test process that allowed the direct observation of acquiring knowledge about the problem space under controlled circumstances (Funke, 2001; Kröner, 2001). In early CPS tests, the problem solvers opportunities to fully explore the problem space by trying out different strategies was limited because the performance scores were influenced by prior (explorative) actions. In contrast, implementing a free exploration phase in modern CPS tests opened up the possibility to separate between the acquisition of knowledge about the underling problem space and the performance when reaching given targets. Thus, modern CPS tests such as MultiFlux or MircoDYN separate the two processes of knowledge acquisition and knowledge application.

Knowledge acquisition refers to the process of gathering knowledge about a non-transparent and potentially dynamically changing system and includes the exploration of the unknown problem space via targeted interactions between the problem solver and the problem situation (Funke, 2001). Knowledge application refers to the process of applying the previously gathered knowledge to reach a given target state or, put simply, to solve the problem (Novick & Bassok, 2005). As mentioned, contemporary CPS tests usually assess these two processes in distinct phases and, accordingly, report separate scores for them. Thereby, both processes correlate substantially, even after controlling for reasoning (Wüstenberg et al., 2012).

Most of the previously mentioned CPS tests are considered one-item-tests in which the variables as well as the relations between them remain constant throughout the testing session. Thus, all indicators assessing knowledge acquisition and knowledge application stem from the very same system configuration and, consequently, depend on each other. In contrast, the use of multiple items, which differ from each other in the variables and the relations between them, resolved this issue. Thereby, this approach contributed to a proper assessment of CPS abilities, for example, by allowing to calculate adequate estimates of reliability (Greiff, Fischer, Stadler, & Wüstenberg, 2015). Therefore, current and widely-used CPS assessment tools such as MicroDYN are based on the so called multiple complex systems approach (MCS, also referred to as multiple item approach; Greiff et al., 2012; Wüstenberg et al., 2012).

Within the multiple complex systems approach, participants work subsequently on several CPS tasks. Within a single CPS task, the structure is based on linear structural equations (Funke, 2001). Thus, every task contains a set of input variables and a set of output variables that are related to each other by linear equations (see Fig. 2.2). Across different CPS tasks, the number of input and output variables as well as the number and type of relations between them can be varied. With regard to the relation types, direct and indirect effects can be distinguished. Direct effects, also referred to as non-dynamic effects, represent the relations between input and output variables (e.g., the relation between input variable B and output variable Y in Fig. 2.2). Indirect effects, also referred to as dynamic effects, represent the relations of output variables among themselves (e.g., the effect of output variable X on itself in Fig. 2.2). These effects lead to dynamic changes of the problem situation irrespective of external interventions (Greiff, Niepel, Scherer, & Martin, 2016).

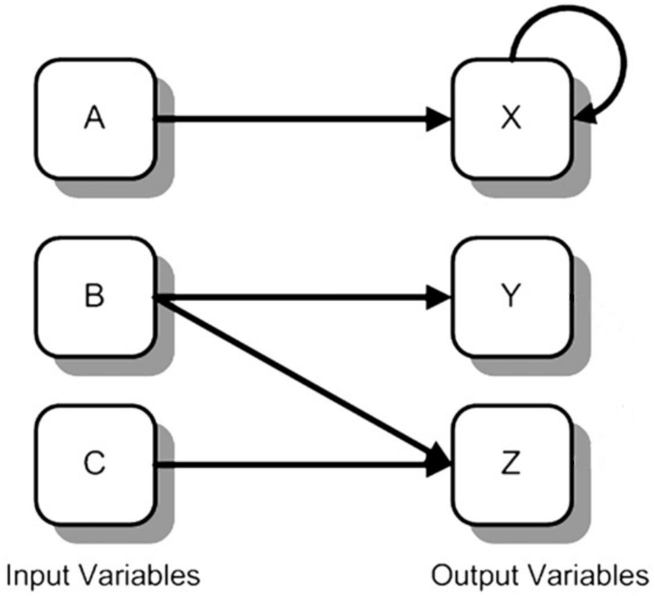


Figure 2.2: Structure of a typical CPS task within the multiple complex system approach (according to Greiff et al., 2012), displaying three input variables (A, B, and C) and three output variables (X, Y, and Z). The arrows between the input and the output variables represent direct, non-dynamic effects. The arrows that output variables have on themselves represent the indirect, dynamic effects.

Among the modern CPS tests, MicroDYN is one of the most frequently used CPS assessment tool because it contains the before mentioned features that establish good prerequisites for a proper and psychometrically-sound assessment of CPS (i.e., no confoundation with prior knowledge, evaluation-free exploration phase, multiple tasks; Greiff et al., 2012; Wüstenberg et al., 2012). Typically, seven to nine fully independent tasks with different cover stories such as feeding a cat, training a handball team, or providing medical aid are implemented in MicroDYN. Variables are always labeled without deep semantic meaning or entirely fictitious in order to avoid the influence of background knowledge. For instance, in the task ”Medical Aid” (see Fig. 2.3) different pharmaceuticals were labeled fictitiously as ”Sarol”, ”Rexol”, and ”Menol”. During the knowledge acquisition phase of MicroDYN, students can freely explore the relations between the pharmaceuticals (left part of Fig. 2.3) and some human health characteristics (right part of Fig. 2.3) by manipulating the input variables. To document the knowledge that was acquired by exploring the unknown problem space, the causal model has to be drawn (lower part of Fig. 2.3). During the knowledge application phase of

MicroDYN, students are provided the correct causal model, independently of the correctness of the previously drawn model. Using the displayed correct model, given target values have to be reached.

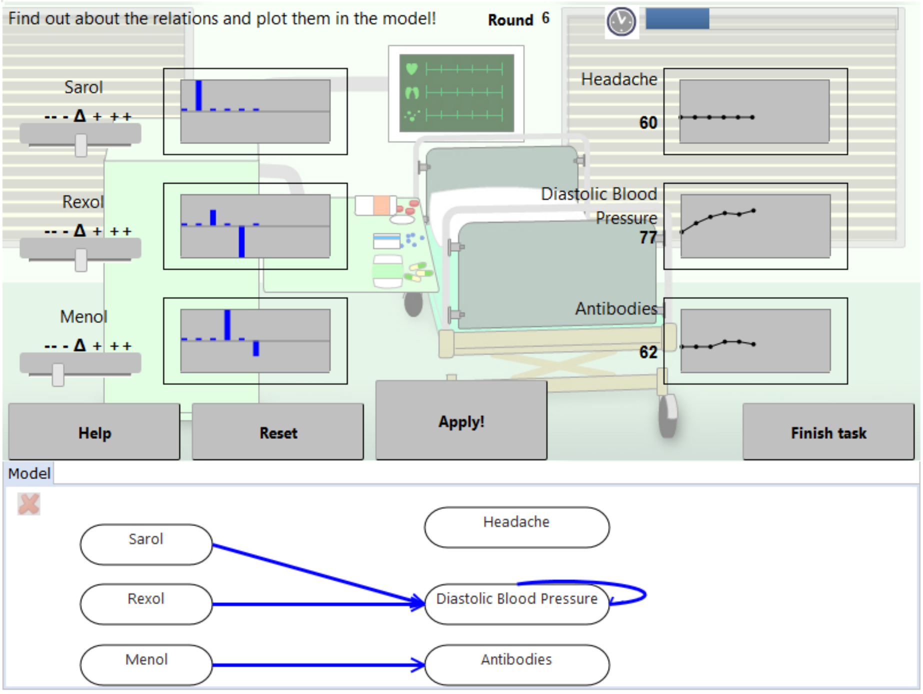

Find out about the relations and plot them in the model!

Figure 2.3: Screenshot of the MicroDYN task ”Medical Aid” during the knowledge acquisition phase (according to Lotz et al., 2017). Participants had to find out the relations between three input variables (pharmaceuticals: Sarol, Rexol, and Menol) and three output variables (human health characteristics: Headache, Diastolic Blood Pressure, and Antibodies) and to plot them in the model.

In accordance with the two phases knowledge acquisition and knowledge application, MicroDYN provides performance data about how successful students performed in those two phases. In addition, process data are also available due to the computer-based test administration. Assessing process data has the advantage of not only obtaining final outcomes, but also of documenting the problem solver's steps towards the specific outcome. Gathering data on every single action of a problem solver provides further insights in applied strategies and tactics or committed errors and could help to understand the problem solving process. Thus, in the following sections, CPS performance data and CPS process data are considered separately to provide a more comprehensive picture about the relations between CPS and intelligence, as well as the incremental validity of CPS for educational success beyond intelligence.

2.2.2 Relation between Performance Measures of Complex Problem Solving and Intelligence

CPS and intelligence share a high theoretical overlap but also exhibit differences. For example, intelligence tests and complex problem solving tests both require the analysis of variables and their relations to each other. Moreover, both require the identification of rules to describe the change in variables. In contrast, complex problem solving tests contain features that are not part of conventional intelligence tests. For instance, the solving of complex problems requires deriving goals, actively searching for relevant information, or selecting actions that contribute to the solution of the problem (for an overview see Funke, 2006). As a consequence, the controversial whether CPS is distinct to intelligence or just a new label for established cognitive constructs such as reasoning emerged and is still not resolved (e.g., Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016; Sonnleitner, Keller, Martin, & Brunner, 2013; Stadler, Becker, Gödker, Leutner, & Greiff, 2015).

On the one hand, there are theoretical arguments that point towards the unique aspects of CPS. Being a successful problem solver requires cognitive-intellectual processes such as multi-step planning, generating adequate mental models, translating mental models into specific plans, and carrying out these plans. Although these cognitive demands are frequently considered as part of the definition of intelligence, they are virtually never found in conventional intelligence tests (Wüstenberg et al., 2012). In comparison to cognition types such as reasoning and mental speed (that are adequately covered by conventional intelligence tests; Raven, 2000; Wüstenberg et al., 2012), CPS requires the problem solver to apply more complex cognition such as dynamic interactions with the task. Because conventional intelligence tests are static and not interactive, these aspects are not covered (Funke, 2010).

Accordingly, first empirical research on CPS, using the CPS tests Lohhausen or Tailorshop revealed relations around zero between intelligence and CPS, which were interpreted as a support for the notion that intelligence and CPS are mostly independent constructs (e.g., Dörner et al., 1983; Putz-Osterloh, 1981, 1985; Putz-Osterloh & Lüer, 1981). This led to the conclusion that a new form of intelligence was discovered, the operative intelligence (Dörner, 1986). However, these mentioned early studies, pointing towards the distinctness of CPS and intelligence, incorporated some methodological weaknesses. For example, the administered CPS tests suffered from reliability problems and the examined samples were highly homogeneous which caused restricted variances on the variables and, in turn, led to lowered relations (Funke, 1983, 1984). Improving these features by examining a more heterogeneous sample and using an improved method to score the Tailorshop performance yielded substantial relations between CPS and intelligence: intelligence explained 28% of the variance of the Tailorshop performance (Funke, 1983). Comparable substantial correlations between intelligence and CPS were received for other refined classical CPS tests such as another improved version of the Tailorshop, Power Plant, and LEARN ($.35 \leq r \leq .46$; Süß, 1999).

Moreover, the relation between intelligence and performance in classical CPS tests might be confounded with prior knowledge (Leutner, 1992, 2002). Thus, modern CPS tests, as used in current research, revealed substantial correlations with intelligence (Stadler et al., 2015). As these correlations were smaller than 1, they were sometimes interpreted as an indicator for the distinctiveness of CPS of intelligence (e.g., Greiff, Wüstenberg et al., 2013; Wüstenberg et al., 2012). However, these conclusions might not be justified. Because correlations smaller than 1 do not necessarily imply the presence of distinct dimensions, it is recommended to assure such claims by further analyses such as higher-order modeling, complemented by providing omegaHS (an effect size of unique latent variable strength) to uncover whether specific dimensions really exist (Gignac & Kretzschmar, 2017).

Contrarily to the distinctiveness hypothesis, there are strong arguments for the assumption that CPS is a subfacet of intelligence. The term 'problem solving' is part of almost every definition of intelligence (Gottfredson, 1997; Rost, 2013) and, accordingly, a meta-analysis reported a substantial correlation between CPS and intelligence of M(g) = .43 (Stadler et al., 2015). Reviewing the relevant studies in more detail, there are hints that the correlation between CPS and intelligence depends on the used assessment methods: Classical CPS tests and narrowly assessed intelligence correlate rather weakly, whereas modern CPS tests and broadly assessed intelligence correlate very highly. A corresponding moderator analysis of the interaction of the assessment methods (CPS: classical vs. modern; intelligence: narrow vs. broad) in the meta-analysis of Stadler et al. (2015) revealed to be non-significant for the combination of classical CPS test and either narrowly or broadly assessed intelligence. However, the combination of modern (and more reliable) CPS measures and either broadly or narrowly assessed intelligence remained unconsidered in the meta-analysis because of too few studies that fit in those categories.

Descriptively reviewing the recent studies that employed modern CPS tests and rather narrowly indicated measures of intelligence (mostly only figural reasoning), typically medium high correlations in the area of approximately $.40 \leq r \leq .60$ were reported (see Table 2.1 for details). In contrast, the very few studies that assessed both modern measures of CPS and broadly indicated intelligence reported considerably higher correlations that ranged in the area of *r* = .65 (adjusted *r* = .75; Kröner, Plass, & Leutner, 2005) up to *r* = .86 (Danner, Hagemann, Schankin, Hager, & Funke, 2011). Thus, the relation between CPS and intelligence seems to be dependent on their assessment methods as a broader indication of the *g*-factor appears to go in hand with higher CPS-intelligence-correlations (see Table 2.1 for details). Moreover, correlations of this magnitude

($.65 \leq r \leq .86$) are typically found among conventional intelligence tests (e.g., Valerius & Sparfeldt, 2014). A theoretically conceivable reason for increasing correlations as a consequence of broadening the intelligence assessment might be that the overlap between CPS and intelligence is not only limited to the reasoning facet. Other aspects of intelligence such as mental speed might also be relevant for successful problem solving if, for example, time restrictions exist (Süß, 1996).

Taken together, there is no consensus about the exact location of CPS in the nomological network of intelligence. It is theoretically conceivable that CPS exhibits unique aspects that are not shared with *g*, but there is also support for the assumption that CPS might be a subfacet of intelligence. However, because the correlation between CPS performance measures and broadly assessed intelligence is very high, it appears unlikely that CPS could exhibit large unique aspects beyond *g*. Nevertheless, if CPS features such unique effects, they should predict real-life criteria as, for example, educational success.

Thus, it appears to be an interesting research goal to examine the relation between CPS and broadly operationalized intelligence in an independent sample. Furthermore, it is one goal of this dissertation project to examine whether CPS performance measures are able to explain additional variance in educational success beyond the established construct of intelligence in spite of the assumingly high correlation between CPS and *g*.

2.2.3 Performance Measures of Complex Problem Solving as a Predictor of Educational Success

The demands of the two CPS facets knowledge acquisition and knowledge application (see section 2.2.1) resemble some of the demands students are confronted with within school contexts. Across several school subjects, students are frequently requested to identify causal relations, to gain knowledge about unknown systems, and to apply this knowledge in new situations and contexts. Thus, it seems worth examining whether the assumed unique aspects of CPS incrementally predict educational success beyond intelligence.

Prior research concerning this topic revealed promising but partly inconsistent results. On the one hand, several studies claimed that CPS is capable of measuring higher-order thinking skills that are not tapped by conventional intelligence tests and, therefore, that CPS is predictive for educational success beyond reasoning (Greiff, Fischer et al., 2013; Greiff, Wüstenberg et al., 2013; Kretzschmar et al., 2014; Wüstenberg et al., 2012). In these studies, CPS (measured by MicroDYN) explained up to 6% of additional variance beyond reasoning (mostly measured by Raven's Matrices [APM, Raven, 1958] or CFT [Culture Fair Test 20-R, Weiß,

Table 2.1: Relations between modern CPS tests and either narrowly indicated intelligence in the sense of (figural) reasoning or broadly indicated intelligence in the sense of a good *g*.

	CPS test	CPS measure	Intelligence measure	Coefficients
Relations between modern CPS tests and narrowly assessed intelligence				
Greiff, Fischer et al., 2013	MicroDYN	Overall performance	Fig. reasoning	r = .50
Greiff, Fischer et al., 2015	MicroDYN	Overall performance	Fig. reasoning	r = .48
	MicroFIN	Overall performance		r = .53
	Genetics Lab	Overall performance		r = .56
	Taylorshop	Overall performance		r = .24
Greiff, Wüstenberg et al., 2013	MicroDYN	Knowl. acquisition	Fig. reasoning	β = .52
		Knowl. application		β = .50
Greiff & Neubert, 2014	MicroDYN	Overall performance	Fig. reasoning	r = .48
Kretzschmar, Neubert, & Greiff, 2014	MicroDYN	Overall performance	Ver. reasoning	β = .58
Kröner, 2001	MultiFlux	Knowl. acquisition	Fig. reasoning	r = .37
		Knowl. application		r = .30
Neubert, Kretzschmar, Wüstenberg, & Greiff, 2015	MicroDYN	Knowl. acquisition	Fig. reasoning	r = .63
		Knowl. application		r = .60
Sonnleitner et al., 2012	Genetics Lab	Overall performance	Fig. reasoning	r = .27
		Overall performance	Num. reasoning	r = .34
Sonnleitner et al., 2013	Genetics Lab	Overall performance	Fig. and num. reasoning	r = .62
Wüstenberg et al., 2012	MicroDYN	Knowl. acquisition	Fig. reasoning	r = .63
		Knowl. application		r = .56
Relations between modern CPS tests and broadly assessed intelligence				
Danner et al., 2011	MicroDYN	Overall performance	General Intelligence	r = .86
Kretzschmar et al., 2016	MicroDYN	Overall performance	Fig., ver., and num. reasoning	r = .85
Kröner et al., 2005	MultiFlux	Knowl. application	Fig., ver., and num. reasoning	r = .75

Note. Knowl. = knowledge; fig. = figural; num = numerical; ver. = verbal

2006]) and showed a higher relevance for the mathematics compared to the verbal domain. Thus, most findings supported the importance of CPS above and beyond reasoning in educational contexts. On the other hand, the study of Sonnleitner et al. (2013) revealed partly contradicting results. CPS (measured by Genetics Lab; Sonnleitner et al., 2012) only demonstrated incremental validity beyond reasoning if the indicators of educational success were computer-based, but not if they were paper-and-pencil-based. The authors concluded that the additionally explained variance might be attributed to the mode of test administration and, thus, that CPS revealed negligible incremental effects for educational outcomes beyond conventional intelligence tests.

Besides these pattern of results that are calling for further research, all of the above mentioned studies share one common characteristic: Intelligence was operationalized by just figural reasoning or by just a few reasoning subtests. Therefore, the intelligence assessments of the mentioned prior studies did not meet the outlined criteria of a good *g* (see section 2.1.1). Relying on those criteria, the studies used reliable but rather few subtests ($<$ 10 subtests). Regarding the diversity of the subtests, the BIS is well-suited as a scheme to determine the broadness of the intelligence assessment. For example, Raven's Matrices or the CFT could be fitted in the crossing between the content facet figural and the operation facet reasoning. The subtests used by Sonnleitner et al. (2013) fitted into the two BIS-crossings figural reasoning and numerical reasoning. Thus, all mentioned studies assessed only one or two crossings. This indicated rather a uniformity of the group factors instead of diversity (see Fig. 2.4). Thus, the above mentioned studies did not fulfill the criteria of a good *g* and this might affect the interpretation of the studies' results. As the correlation between intelligence and CPS seems to increase with a broader operationalization of intelligence, the probability of CPS exhibiting substantial unique aspects might decrease and, therefore, limit the incremental validity of CPS beyond intelligence.

In fact, only one recent study investigated the incremental validity of CPS for educational success beyond broad intelligence measures (Kretzschmar et al., 2016). When indicating intelligence in the sense of a good *g*, no incremental effects of CPS were found beyond intelligence for GPA (ΔR^2 = .007). In contrast, when indicating intelligence by the common narrow operationalization in the sense of figural reasoning, CPS additionally explained 3.9% of the variance in GPA which was comparable to prior studies. These results turned out as theoretically expected and were encouraging to further investigate the incremental validity of CPS for educational outcomes beyond broadly assessed intelligence. The study of Kretzschmar et al. (2016) as well as almost all CPS studies in educational context only used school grades or GPA as indicators of student's scholastic achievement. Although assessing grades has several advantages such as their easy accessibility and their predictive validity over long periods of time, grades could also be

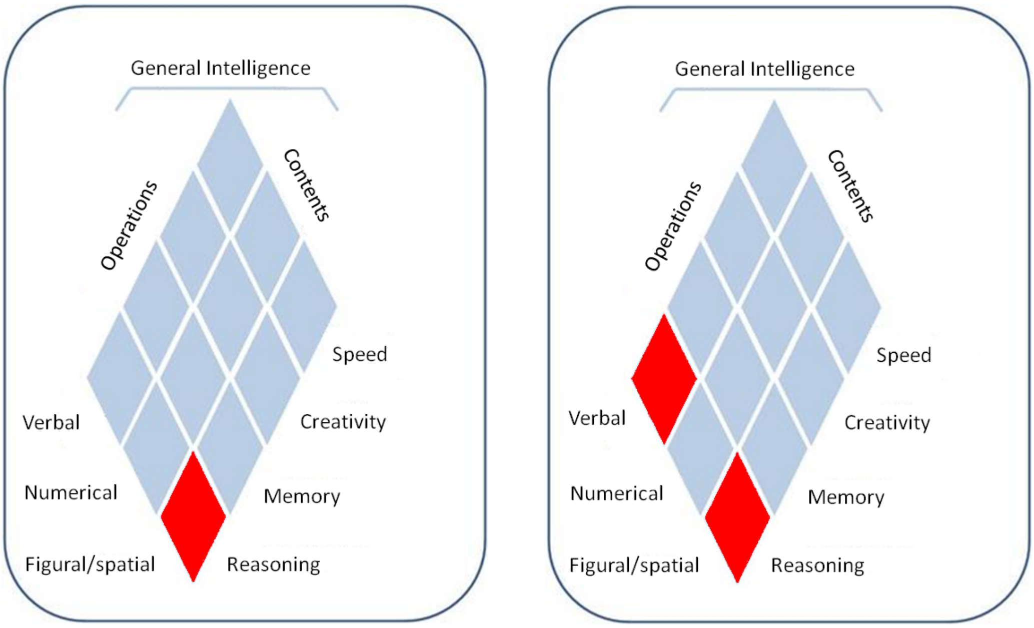


Figure 2.4: Intelligence assessments of most prior studies on the incremental validity of CPS. Left side: narrow operationalization of intelligence in the prior studies that used Raven's Matrices or Culture Fair Test (CFT 20-R); right side: narrow operationalization of intelligence by two reasoning subtests as assessed in Sonnleitner et al. (2013). Covered cross-classifications are marked in red.

criticized for several reasons. For instance, teachers' judgments might be biased and, thus, differ systematically from school to school or even from class to class. The application of a wider range of outcome variables by assessing students' competences with school grades and with scholastic competence tests as a more standardized and objective measure of educational success might overcome the restrictions accompanied by only relying on school grades.

Taken together, several prior studies investigated the incremental validity of CPS for educational outcomes beyond measures of intelligence. However, in those studies, intelligence was assessed either just as figural reasoning or with only a few reasoning subtests and not in the sense of a good *g*. As a consequence, CPS was probably less correlated with measures of intelligence and, thereby, the probability that CPS explained additional variance in the criteria beyond a "weak *g*" was increased. Moreover, assessing mostly only GPA as the relevant criterion might also restrict the interpretation of the results. In turn, the assessment of grades and scholastic competence tests would increase the informative value. In addition, assessing educational outcomes in school subjects of different domains gives the opportunity to detect differential result patterns.

Thus, it is one research aim of this dissertation project to examine the incremental validity of CPS beyond broadly assessed intelligence for educational success measured by grades and scholastic competence tests in mathematics and German as the two core school subjects.

2.2.4 Process Measures of Complex Problem Solving and their Relation to Intelligence

The computer-based assessment of CPS has many advantages compared to conventional intelligence tests. Besides final CPS performance measures, CPS assessment tools provide the possibility to analyze every single step the problem solver takes towards the solution of the problem by documenting every single action in computer-generated log-files. For instance, direct and detailed information about how a problem solver might use strategies and tactics or commits errors while exploring an unknown problem space can be derived. Although past research repeatedly praised this potential, it was, in fact, only seldom utilized. Prior studies mainly took advantage of the potential of CPS process measures by examining time-on-task (e.g., Goldhammer et al., 2014; Scherer, Greiff, & Hautamäki, 2015) or by investigating rather specific strategies that were strongly dependent on the particular CPS test (e.g., Güss, Tuason, & Orduña, 2015; Strohschneider & Güss, 1999). Only very few studies examined the usage of domain-general effective strategic behaviors such as VOTAT or NOTAT (Greiff et al., 2016; Greiff, Wüstenberg, & Avvisati, 2015; Wüstenberg, Stadler, Hautamäki, & Greiff, 2014).

Typically, when exploring the problem space of CPS tasks, direct effects and dynamic effects might occur and have to be identified (see section 2.2.1). Accordingly, applying the strategic behaviors VOTAT (Tschirgi, 1980; also known as ”control-of-variables strategy” [CVS], Chen & Klahr, 1999) and NOTAT (Lotz et al., 2017; also known as ”non-interfering observations”, Greiff et al., 2016) are effective for detecting these effect types.

Referring to VOTAT, this strategic behavior requires to systematically vary one variable at a time while all other variables remain constant to isolate the one effective variable that is responsible for a particular effect. Its significance is based on its domain-general applicability and, therefore, VOTAT is regarded as the key strategy in scientific reasoning. Nevertheless, most students have no generalized understanding of VOTAT because it does not routinely develop without practice (Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016; Zimmerman & Croker, 2013). The acquisition of knowledge about VOTAT is rather a progressive process: students do not simply abandon invalid strategies once they discovered valid ones, but they use a mixture of former invalid and newer valid strategies while the application of the new optimal strategy is progressively increasing (Chen & Klahr, 1999; Schauble, 1996; Vollmeyer, Burns, & Holyoak, 1996). Specifically, in the context of CPS, the application of VOTAT leads to more successful performance in CPS knowledge acquisition as well as knowledge application, as it is the most effective strategic behavior to identify direct effects between input and output variables (Greiff et al., 2016; Greiff, Wüstenberg, & Avvisati, 2015; Kröner et al., 2005; Wüstenberg, Stadler et al., 2014).

Referring to NOTAT, this strategic behavior is characterized by systematically constraining all variables to simultaneously remain at a zero level. In scenarios in which, for example, dynamic growth or decay effects have to be identified, the NOTAT strategy is optimal because the problem solver can actively observe how the system is changing itself independently of the problem solver's manipulations. In CPS contexts, problem solvers typically act too quickly because they struggle with resisting the temptation to manipulate the variables immediately. In contrast, proficient problem solvers monitor the autonomously changing system and, thus, are more successful in solving dynamic CPS tasks (Dörner, 1980b; Dörner & Schaub, 1994). In recent CPS research, NOTAT's significance was demonstrated by its unique effects on CPS performance after controlling for VOTAT (Greiff et al., 2016).

Reviewing the above mentioned empirical studies on strategic CPS behaviors, they exhibited several methodological issues. First, previous studies analyzed the strategic behaviors either averaged across all administered tasks or focused only on one single task. Second, almost all prior studies investigated the VOTAT or NOTAT behaviors dichotomously as credit was given if students applied the particular strategic behavior at least once for each input variable in case of VOTAT or at least once at all in case of NOTAT. Third, yet no study that had considered NOTAT differentiated between dynamic tasks (for which applying NOTAT is effective) and non-dynamic tasks (for which applying NOTAT is not effective).

With regard to these methodological issues, previous studies left several questions unanswered. Therefore, it is a research aim of this dissertation project to, first, display the temporal course of the strategic behaviors across a task set by separately analyzing every administered task. Second, to provide more profound insights in the application frequencies of the strategic behaviors and how those might change across the task set by using a continuous scoring of VOTAT and NOTAT in the manner of relative frequencies (Kröner et al., 2005). Third, to reveal a more conclusive picture about the importance of NOTAT for dynamic tasks and, thereby, to deepen the understanding of students' flexibility in adapting effective strategic behaviors when confronted with changing task types by distinguishing between dynamic and non-dynamic tasks.

Moreover, the relations between the strategic behaviors (i.e. VOTAT and NOTAT) and intelligence are also widely unexamined. As mentioned in section 2.2.2, there is a substantial and positive correlation between intelligence and

CPS performance measures. Whereas this relation provides a rather general perspective, the question arises to what extend intelligence and specific CPS behaviors are linked. Empirical studies on the relation between intelligence and VOTAT in the context of CPS as well as scientific reasoning suggested that both are substantially positively connected (see Table 2.2 for details). Thus, the ability to select and apply VOTAT could be regarded as ”a manifestation of intellectual ability” (Veenman, Wilhelm, & Beishuizen, 2004, p. 91).

As an extension of this line of argumentation, not only the application but also the adaption of effective strategic behaviors when confronted with changing task demands might be considered as an aspect of intelligence. Accordingly, more intelligent students should be able to learn harder things more quickly (Thorndike, 1922). Guthke and Stein (1996) adopted this proposal and showed that the performance on learning and intelligence tests were highly related (r = .83). Moreover, higher levels of intelligence are also related to adapting problem solving strategies to changing situations more effectively (Benedek, Jauk, Sommer, Arendasy, & Neubauer, 2014; Deák, 2003; LePine, Colquitt, & Erez, 2000). Thus, recognizing that a strategic behavior is effective to solve a given problem and, therefore, applying and adapting it indicates intelligence.

Taken together, past research on strategic CPS behavior shared methodological issues that prevented them from providing a conclusive picture about the temporal course of the strategic CPS behaviors VOTAT and NOTAT across a task set. Thus, it is a goal of this dissertation project to fill this research gap by applying more adequate research methods such as analyzing relative frequencies and distinguishing between different task types. Moreover, as another research aim of this dissertation project, the understanding about the role of intelligence in the application and adaption of VOTAT and NOTAT will be deepened. Thus, their relation is examined across a task set with changing demands to give further insights in how intelligence takes action during the problem solving process.

Table 2.2: Overview of studies on the relation between intelligence and the use of VOTAT in the domains of CPS and scientific reasoning.

	Test instrument	VOTAT operationalization	Intelligence measure	Correlation	Sample size	Participants' grade level and mean age
Domain CPS						
Kröner et al., 2005	MultiFlux	Relative frequency	Figural, verbal, and numerical reasoning	r = .41	101	Grades: 9-12; *Age*: M = 15.0, SD = 1.0
Wüstenberg, Stadler et al., 2014	MicroDYN	Dichotomous per task (VOTAT application at least once for every input variable)	Verbal and numerical reasoning	r = .64	3191	Grades: 6 and 9; *Age*: M = 13.59, SD = 1.56
Domain scientific reasoning						
van der Graaf, Segers, & Verhoeven, 2015	Ramp task	(a) Absolute frequency of correctly designed experiments after two trials each	Nonverbal reasoning	r_a = .42	46	Age Range: 4;6-6;3, M = 5;3
		(b) Absolute frequency of correctly set variables		r_b = .47		
Künsting, Kempf, & Wirth, 2013	learning environment	Relative frequency	Figural reasoning	r = .30	129	Grade: 9; *Age*: M = 14.33, SD = 0.69
Veenman, Bavelaar, De Wolf, & Van Haaren, 2014	Otter task	Absolute frequency	Figural, verbal, and numerical reasoning	r = .17	52	Grade: 7; *Age*: M = 13;2

Note. Table taken from Lotz et al. (2017)

2.2.5 Process Measures of Complex Problem Solving as Predictors of Educational Success

In educational contexts, scientific reasoning strategies such as VOTAT are fundamental for mastering school science and mathematics (Bitner, 1991; Bryant, Nunes, Hillier, Gilroy, & Barros, 2015; Tajudin & Chinnappan, 2015). Reviewing the demands of scientific reasoning and of mathematics problem solving, they tend to exhibit high theoretical overlap. For instance, mathematics problem solving contains a broad range of thinking skills such as recognizing patterns, identifying concepts and relations, making inferences from data, challenging results, or modifying conclusions (OECD, 2014; Schoenfeld, 2014). Comparably, scientific reasoning encompasses generating hypotheses, acquiring knowledge by strategic actions, testing theories, deriving conclusion, or revising knowledge (Klahr & Dunbar, 1988; Morris, Croker, Masnick, & Zimmerman, 2012). Also empirically, scientific reasoning and mathematics problem solving overlap substantially positive (e.g., Bitner, 1991: $r = .59$). Given this connectedness between scientific reasoning and mathematics problem solving, it could be expected that students with higher levels of scientific reasoning skills – and especially high VOTAT skills – might also obtain higher scholastic achievements in mathematics or science education.

Accordingly, empirical studies highlighted VOTAT's relevance for mathematics (Bitner, 1991; Tajudin & Chinnappan, 2015) and science education (Adamson et al., 2003; Bitner, 1991; Bryant et al., 2015; Huppert, Michal-Lomask, & Lazarowitz, 2002). For the school subject mathematics, it was shown that students, who had a better understanding of the principles of scientific reasoning, scored higher in a high-stakes mathematics achievement test (Tajudin & Chinnappan, 2015). Comparably, in the study of Bitner (1991), the proficient use of VOTAT explained 12% of the variance of a composite mathematics grades score (consisting of grades in five different mathematics courses) and 15% of the variance of a composite science grades score (consisting of grades in five different science courses). Additionally, studies on the importance of VOTAT revealed comparable effects for the school subject biology (Adamson et al., 2003; Huppert et al., 2002).

Besides this promising evidence about the importance of VOTAT for success in mathematics and science, the studies mentioned above shared one characteristic that limits the interpretation of the outlined results: it was not controlled for the common variance with intelligence. Yet, only one study considered this issue and revealed that even after controlling for intelligence, the ability of producing conclusive variable manipulations, i.e. applying VOTAT, longitudinally predicted learning progress in science education over a three year period (Bryant et al., 2015). However, in this study, students worked on a control-of-variables task that is often referred to as 'the inclined plane problem'. Unfortunately, tasks like this were sometimes used as indicators of reasoning (see Kretzschmar et al., 2014), which might have biased the results. Thus, an operationalization of VOTAT that is closer to actual behavior might grant more valid insights into the interplay of intelligence, strategic behaviors, and educational success. Process measures about the strategic VOTAT behaviors during the exploration phase of CPS tasks might represent exactly such behavior-based VOTAT indicators.

Moreover, in the context of CPS, examining the incremental validity of the strategic VOTAT and NOTAT behaviors for educational success beyond intelligence might represent the next crucial step in research on CPS. As mentioned in section 2.2.2, CPS tasks exhibit unique features, such as being dynamic and non-transparent, that are not assessed by conventional intelligence tests. As another major difference, CPS-tasks additionally contain an exploration phase in which the problem solver actively interacts with the task by applying strategic behaviors. Capturing the active exploration behavior that is manifested in the application of VOTAT and NOTAT might be the key to understand the assumed unique aspects of CPS that might be incrementally predictive for educational success beyond intelligence.

Accordingly, it is another research aim of this dissertation project to examine whether the strategic CPS behaviors VOTAT and NOTAT are capable of predicting educational success above and beyond intelligence.

2.3 Motivation and Educational Success

There is ample evidence that motivation functions as an important determinant for educational success (e.g., Hattie, 2009; Valentine et al., 2004). Thus, motivational constructs possibly have the potential to account for the considerable amounts of variance in students' academic achievement that are left unexplained by general intelligence. Accordingly, various motivational constructs such as academic self-concepts, academic interest, learning goals, work avoidance, and achievement motives were frequently shown to predict educational success even beyond intelligence (e.g., Chamorro-Premuzic & Furnham, 2008; Freudenthaler, Spinath, & Neubauer, 2008; Kriegbaum, Jansen, & Spinath, 2015; Meece, Wigfield, & Eccles, 1990; Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013; Schaefer & McDermott, 1999; Spence, Pred, & Helmreich, 1989; Spinath, Spinath, Harlaar, & Plomin, 2006; Zuffianò et al., 2013).

Among the variety of motivational constructs, academic self-concepts and interests appear to be the key aspects of motivation within the well-established expectancy-value theory (Eccles et al., 1983; Wigfield & Eccles, 2000). As both constructs correlate only modestly with intelligence (Helmke, 1992; Jansen et al., 2016), it seems promising to investigate their incremental validity for educational success beyond intelligence. Thus, the following section illustrates the two constructs self-concept and interest within the expectancy-value theory in more detail and reviews their incremental validity for educational success. Thereby, the differential relevance of motivational and cognitive variables for different indicators of scholastic achievement is highlighted. Specifically, it seems that intelligence is the better predictor for scholastic competence tests, whereas motivation seems to be the better predictor for grades.

2.3.1 Self-concept and Interest as Predictors of Educational Success

According to expectancy-value theory (Eccles et al., 1983; Wigfield & Eccles, 2000), the expectancy to succeed in a task and the value assigned to the task determine achievement-related behavior. With regard to the expectancy component, the model posits that the expectancy for success, also referred to as the self-concept, is influenced by the individuals' perception about how well an upcoming task will be done based on their current competences. Concerning the value component, the model specified the intrinsic value of a task, also referred to as interest, as the extent to which a person gains enjoyment from doing the task and not for anticipated consequences. Typically, academic self-concepts and academic interests are domain-specifically structured (Marsh, Smith, Barnes, & Butler, 1983). They tend to correlate only slightly across different domains, but they are highly related within one domain (Bong & Clark, 1999; Rost & Sparfeldt, 2002). For instance, in the school subject mathematics, correlations between self-concept and interest reached values of $r = .74$ (Rost, Sparfeldt, & Schilling, 2007) or $r = .80$ (Trautwein et al., 2012) for German high school students.

Reviewing the relations of both components with educational success, academic self-concepts repeatedly exhibited moderate relations of about $.30 \leq r \leq .60$ to scholastic achievement (Guay, Marsh, & Boivin, 2003; Valentine et al., 2004), whereas academic interests typically only revealed weak to moderate relations to achievement ($.20 \leq r \leq .30$; Jansen et al., 2016; Schiefele, Krapp, & Winteler, 1992). This pattern of results is in line with expectancy-value theory: whereas self-concepts seem to be more closely associated with school performance, interests are more predictive for achievement-related choices or efforts (Eccles et al., 1983; Wigfield & Eccles, 2000).

2.3.2 Differential Relevance of Cognitive and Motivational Variables

Reviewing the large body of empirical research about the predictive power of cognitive and motivational variables, past research suggested a differential relevance of motivational and cognitive variables when relying on competence tests or grades as achievement indicators. Intelligence showed a closer relation to scholastic competence tests than to grades as already highlighted in section 2.1.2. In contrast, self-concepts and interests typically revealed a closer relation to grades than to scholastic competence tests (e.g., Helmke, 1992; Jansen et al., 2016; Jansen, Schroeders, & Lüdtke, 2014; Marsh et al., 2005; Möller, Pohlmann, Köller, & Marsh, 2009; Steinmayr & Meißner, 2013; Zaunbauer, Retelsdorf, & Möller, 2009).

Regarding self-concepts, a meta-analysis revealed that school subject-specific self-concepts were numerically closer related to grades than to scholastic competence tests in mathematics and in the verbal domain (mathematics: r_{grades} = .50, r_{test} = .37; verbal: r_{grades} = .40, r_{test} = .34; Möller et al., 2009). For German samples, similar result patterns were obtained in three science school subjects (e.g., physics: β_{grades} = .41, β_{tests} = .11; Jansen et al., 2014). Comparably, school subject-specific interests also revealed numerically higher regression coefficients on grades compared to scholastic competence tests in five school subjects within a German sample (e.g., mathematics: β_{grades} = .42, β_{test} = .23; German: β_{grades} = .18, β_{test} = -.01; Jansen et al., 2016). However, when considering both expectancy-value components at the same time (Marsh et al., 2005), only self-concept revealed numerically higher regression coefficients on mathematics grades compared to a mathematics competence test, whereas interest showed regression coefficients around zero on both criteria. This drop of the formerly substantial interest-achievement-coefficients was probably caused by controlling for the substantial overlap between self-concept and interest (study 1/2: r = .56/.58; Marsh et al., 2005). As mentioned in section 2.3.1, self-concept might be more important than interest for predicting educational success, because choices in school contexts are quite limited (i.e., everybody has to do courses in mathematics or the native language). Further support for the loss of the predictive value of interest after controlling for the shared variance with self-concept was obtained by Meece et al. (1990) and Spinath et al. (2006): School subject-specific interests contributed only neglectably to the prediction of grades in mathematics (Meece et al., 1990) as well as teacher ratings of competences in mathematics, science, and native language (Spinath et al., 2006).

Considering the very few studies that examined the differential relevance of cognitive and motivational variables for scholastic competence tests and grades in concert, an even more conclusive picture emerged. Regarding solely intelligence and self-concept, intelligence accounted for 38% of the variance in a mathematics competence test but only for 20% in mathematics grades, whereas self-concept accounted for only 32% of the variance in the competence test but for 57% in grades (Helmke, 1992). Comparably, when considering solely intelligence and interest (Jansen et al., 2016), intelligence was the better predictor for scholastic competence tests than for grades in five school subjects (e.g., mathematics: $\beta_{\text{test}} = .53$, $\beta_{\text{grades}} = .20$; German: $\beta_{\text{test}} = .41$, $\beta_{\text{grades}} = .17$), whereas interest showed an inverted pattern (e.g., mathematics: $\beta_{\text{test}} = .14$, $\beta_{\text{grades}} = .39$; German: $\beta_{\text{test}} = .04$, $\beta_{\text{grades}} = .20$).

Besides this first evidence for the differential relevance of cognitive and motivational variables, all studies mentioned above reported these differences only numerically, but did not test them statistically. Yet, only the study by Steinmayr and Meißner (2013) statistically compared the path coefficients of intelligence and self-concept on scholastic competence tests and grades in mathematics. Referring to scholastic competence tests, intelligence showed a path coefficient of $\beta = .54$, whereas self-concept exhibited a path coefficient of $\beta = .30$. The difference between both path coefficients revealed to be statistically significant ($p < .001$). However, although the regression coefficients on grades ($\beta_{\text{intelligence}} = .31$; $\beta_{\text{self-concept}} = .53$) turned out to be almost perfectly inverted compared to those on the mathematics competence test, surprisingly, their difference did not reach statistical significance ($p = .96$) This surprising result pattern calls for further examination.

In light of the above reported findings, it is one goal of this dissertation project to examine the differential relevance of intelligence, self-concept, and interest in concert to gain more profound insights into the interplay of these variables when predicting scholastic competence tests and grades. Moreover, statistically testing their differential relevance by more advanced and sophisticated methods warrants a reliable and meaningful testing and interpretation of corresponding path differences. Thereby, a broad assessment of intelligence represents an improvement to prior studies that mostly assessed intelligence in the sense of figural reasoning (e.g., Jansen et al., 2016; Steinmayr & Meißner, 2013) or as a conglomerate of reasoning and other variables (e.g., Helmke, 1992). Moreover, it is another research aim of this dissertation project to evaluate the increments of intelligence, self-concept, and interest beyond each other. As mentioned, the incremental validity of interest might vanish when controlling for the shared variance with self-concept. Thus, the unique variance proportions of each variable when predicting the scholastic competence test and grades are examined in order to avoid an underestimation of the significance of interest as the probably least powerful predictor.

2.4 Research Aims of this Dissertation Project

Based on the theoretical frameworks and empirical findings summarized in the previous sections, we conducted three articles that build the core of this dissertation project. Thereby, we pursued the goal of examining the incremental validity of selected cognitive, strategic-behavioral, and motivational variables for educational success beyond broadly operationalized intelligence. In detail, the three overarching research aims were as follows:

First, we aimed to examine the incremental validity of CPS performance measures beyond broadly operationalized intelligence for educational success.

As a second research goal, we explored the application and adaption of the optimal strategic-behavioral CPS process measures VOTAT and NOTAT, their relations to intelligence, and whether they showed unique effects for educational success beyond comprehensive measures of intelligence.

The third aim was to examine the differential relevance of motivational and cognitive variables, depending on which achievement indicator of educational success was used and whether the predictors incrementally explained variance beyond intelligence and among each other in educational success.

Chapter 3

Empirical Studies

To pursue the aim of examining the incremental validity of selected cognitive, strategic-behavioral, and motivational variables beyond broadly operationalized intelligence on educational success, we conducted three separate empirical studies that formed the basis of this dissertation project. In the following section, the three studies are outlined by shortly summing up their theoretical background and hypotheses, methods, results, and discussions. Important additional findings that were not incorporated in the corresponding publications are provided in supplementary analyses sections.

Study 1 (Lotz, Sparfeldt, & Greiff, 2016) overcame one of the major shortcomings in research on the increment of CPS by broadly operationalizing intelligence instead of assessing only measures of figural reasoning. We investigated whether intelligence was a stronger predictor for educational success than CPS and whether CPS still provides a substantial increment above and beyond broadly assessed intelligence on scholastic competence tests and grades in mathematics and German.

In the second study (Lotz et al., 2017), we explored the application and adaption of two optimal strategic CPS behaviors (i.e. VOTAT and NOTAT) across a set of nine CPS tasks with changing task demands. We inspected the relations between intelligence and the application frequencies as well as the adaption gradients of VOTAT and NOTAT across the task set. In supplementary analyses, maximally efficient behavior was investigated and it was clarified whether VOTAT and NOTAT provided an increment beyond a broad intelligence operationalization on scholastic competence tests and grades.

In study 3 (Lotz, Schneider, & Sparfeldt, 2018), we shifted the focus towards the interplay between intelligence and motivational variables when predicting educational success. We statistically examined the differential relevance of intelligence, self-concept, and interest, depending on which achievement indicator (i.e. scholastic competence tests and grades) was used. By means of advanced statistical methods, we clarified whether intelligence revealed statistically higher coefficients than motivational variables for scholastic competence tests and whether self-concept revealed statistically higher coefficients than intelligence and interest for grades in the school subject mathematics. Moreover, we examined the unique effects of each predictor. Supplementary analyses considered the differential relation pattern for the school subject German.

3.1 Study 1: Lotz, Sparfeldt, & Greiff (2016)

Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex problem solving in educational contexts — Still something beyond a ”good *g*”? *Intelligence, 59,* 127–138.

In this study, we aimed to shed light on the unique aspects of CPS when predicting scholastic competence tests and grades in the core school subjects mathematics and German.

3.1.1 Theoretical Background and Hypotheses

Recent studies concerning the added value of CPS above and beyond intelligence revealed promising first results (e.g., Greiff, Wüstenberg et al., 2013; Wüstenberg et al., 2012): as expected, intelligence and CPS correlated substantially and CPS explained approximately 5 % of the variance in GPA after controlling for intelligence. However, a thorough inspection of these studies revealed that in most cases intelligence was assessed rather narrowly in the sense of figural reasoning. Thereby, the criteria of a ”good *g*” that were outlined in section 2.1.1 (Jensen & Weng, 1994) were not fulfilled. Such a rather narrow operationalization of intelligence entails the risk of lowering the correlation between intelligence and CPS. Moreover, rather narrowly assessed intelligence might only predict rather narrow aspects of the criterion. Thus, the chance for CPS to add substantial amounts of variance beyond intelligence when predicting educational success was increased. Indeed, one recent study overcame this shortcoming and revealed that CPS only predicted GPA beyond the commonly used narrow operationalisation of intelligence, but not if intelligence was assessed broadly (Kretzschmar et al., 2016).

Using GPA as an indicator of educational success has several advantages as its easy accessibility and its predictive validity over longer time periods, but it could also be criticized for some reasons. For example, GPA is a very general indicator of academic achievement and, more generally speaking, grades might be biased as they are based on teacher's judgments and, thus, differ from school to school or from class to class. Therefore, filling the research gap left by Kretzschmar et al. (2016), we assessed (a) specific school grades in the two main school subjects mathematics and German, (b) complemented the specific school grades by scholastic competence tests in mathematics and German, and (c) used a more heterogeneous high school sample (instead of university students) to examine the following three research objectives, using a broad intelligence operationalization that fulfills the criteria of a good *g*:

Hypothesis 1. We expected the correlation between broadly assessed intelligence and CPS to exceed the formerly reported correlations between narrowly assessed intelligence and CPS.

Hypothesis 2. Displaying the correlations of intelligence and CPS with the measures of educational success without controlling for their shared variance, we assumed intelligence and CPS to correlate closely with both criteria of educational success.

Hypothesis 3a. Statistically predicting the measures of educational success and, thereby, controlling for the common variance between intelligence and CPS, we expected intelligence to exhibit higher regression coefficients on the criteria of educational success than CPS.

Hypothesis 3b. Examining the incremental validity of intelligence and CPS for educational success, we expected the intelligence increment to substantially predict educational success beyond CPS and, additionally, we expected the CPS increment to substantially predict educational success beyond intelligence.

3.1.2 Methods

Participants and Procedure. The sample comprised N = 496 German high school students (n = 265 females, n = 3 without gender specification; age: M = 16.39 years, SD = 0.95). All measures were assessed during three consecutive lessons: in the first lesson, students of one class worked together on the intelligence test battery. During the remaining two lessons, students of one class were randomly split in two halves. In one lesson, the first half worked on the reading comprehension test, whereas the second half executed the CPS test. During the remaining lesson, the first half worked on the CPS test, whereas the second half completed the mathematics competence test. Thus, all participants worked on the intelligence test battery and the CPS test, but only half of the sample worked either on the mathematics competence test (mathematics subsample; n = 245) or on the reading comprehension test (German subsample; n = 251).

Instruments. Intelligence was assessed by a selection of 10 subtests of the Berlin Intelligence Structure Test - Form 4 (BIS-4; Jäger et al., 1997) which covered a wide range of the content-operation-combinations. Thus, the intelligence assessment of this study fulfilled the criteria of a good *g* (cf. Jensen & Weng, 1994, see Fig. 3.1; also see section 4.3.1). CPS was assessed with the entirely computer-based microworld program MicroDYN in which the participants actively explored and subsequently controlled an unknown system (Greiff et al., 2012; Wüstenberg et al., 2012). Selected items from the standardized mathematics and reading comprehension competence tests of a German longitudinal program (KESS 10/11 [Competences and Attitudes of Students from Schools in Hamburg]; Vieluf, Ivanov, & Nikolova, 2011) were used to assess scholastic competences in mathematics and German. Additionally, students provided their midterm report card grades in mathematics and German which were reversely scored for a more meaningful interpretation.

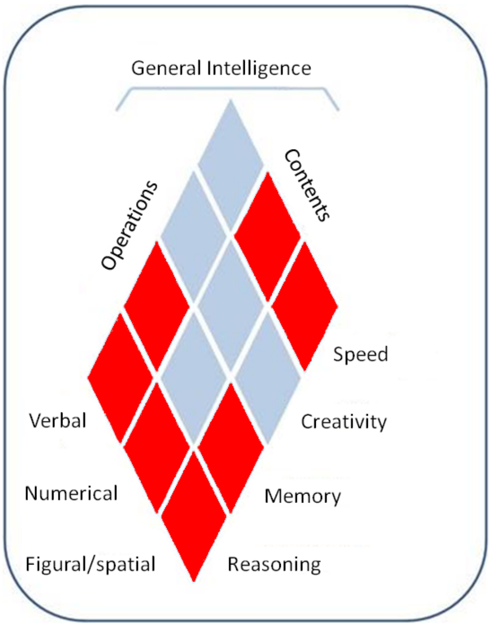


Figure 3.1: Intelligence assessment of this dissertation project, fulfilling the criteria of a good *g*. Covered cross-classifications are marked in red.

Analyses. All analyses were run separately for the two subsamples (mathematics and German subsample). Concerning Hypotheses (1) and (2), two separate latent models with solely bidirectional paths between the latent factors intelligence, CPS, and the scholastic competence test as well as the manifest grade were established. Corresponding correlation coefficients were inspected. Regarding Hypothesis (3a), we established analogous models with unidirectional paths from the predictors intelligence and CPS on the criteria scholastic competence tests and grades. Corresponding regression coefficients were inspected and it was statistically tested whether intelligence was the stronger predictor compared to

CPS for both criteria. With regard to Hypothesis (3b), two models per subsample were specified: one in which CPS and the intelligence-residual (increment which is independent from CPS) were used as predictors and another one in which intelligence and the CPS-residual (increment which is independent from intelligence) were used as predictors. Significant paths from the respective residual factors indicated substantial increments. By squaring the regression coefficients, the percentage of explained variance could conveniently be examined.

3.1.3 Results

As expected in Hypothesis (1), broadly operationalized intelligence and CPS correlated highly in the mathematics subsample (r = .76) as well as in the German subsample (r = .69; see Fig. 3.2). Furthermore, in accordance with Hypothesis (2), intelligence and CPS correlated for the most parts substantially and comparably high with the scholastic competence test in mathematics (r_g = .85; r_{CPS} = .79; p's $<$ .05) and German (r_g = .52; r_{CPS} = .47; p's $<$ .05) as well as with the grades in mathematics (r_g = .42; r_{CPS} = .36; p's $<$ .05). Referring to the grades in German, only intelligence revealed a substantial correlation (r = .23; p $<$ .05), but not CPS (r = .01; p = .45; see Fig. 3.2).

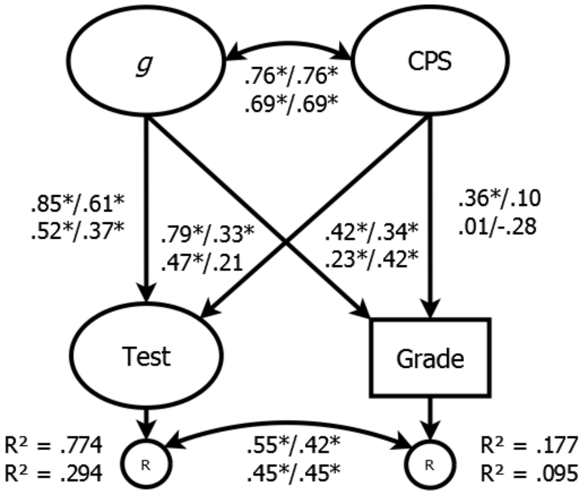


Figure 3.2: Standardized coefficients of the correlation-based/regression-based models of the mathematics subsample in the numbers above and the German subsample in the numbers below. Measurement models were not depicted. R^2 = total percentage of explained variance by intelligence and CPS.
$^*p < .05$.

With regard to Hypothsis (3a), all regression coefficients of intelligence decreased only slightly after controlling for the shared variance with CPS and remained to be substantially. For example, the relation between intelligence and the mathematics competence test changed from r = .85 to β = .61. In contrast, CPS coefficients dropped considerably and reached for the most parts a non-significant level or even became negative (see Fig. 3.2 for further details). Moreover, as expected, intelligence revealed to be the statistically stronger predictor than CPS for the competence tests and grades in both school subjects (all p's $< .05$).

Concerning the incremental validity of the increments (Hypothesis 3b), the intelligence increment significantly added 15.4% (β = .39; $p < .05$) and 7.4% (β = .27; $p < .05$) to the totally explained variance of the competence tests in mathematics and German, respectively. Moreover, intelligence significantly contributed 4.9% (β = .22; $p < .05$) and 9.5% (β = .31; $p < .05$) to the totally explained variance of the grades in mathematics and German, respectively (see left side of Fig. 3.3). In contrast, the CPS residual significantly added only 4% (β = .21; $p < .05$) of explained variance beyond intelligence to the totally explained variance of the mathematics competence test (see right side of Fig. 3.3). Regression coefficients of the CPS residual on the other three criteria were all non-significant (all β $< .15$; all p's $> .14$).

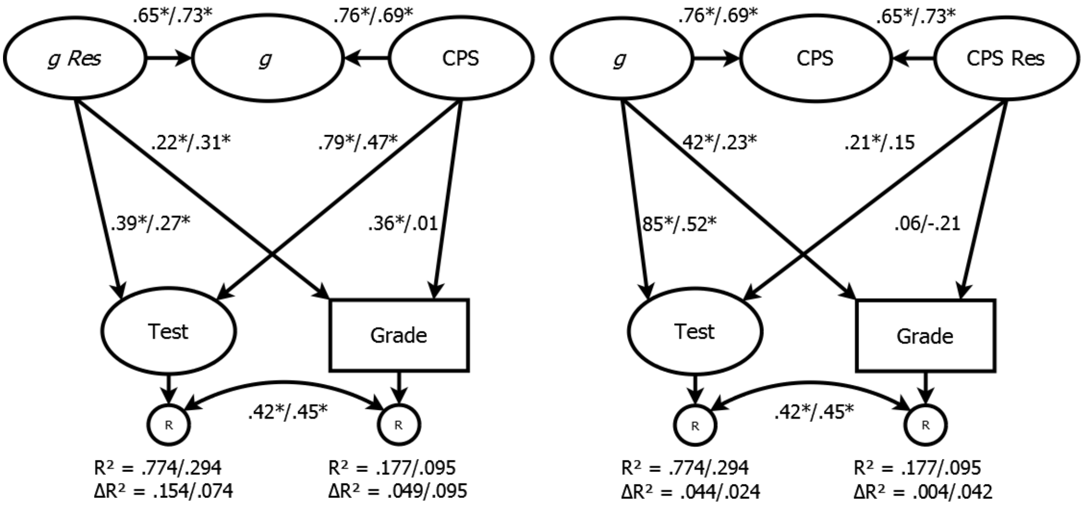


Figure 3.3: Standardized coefficients of the intelligence-residual model of the mathematics/German subsample on the left side and of the CPS-residual model of the mathematics/German subsample on the right side. Measurement models were not depicted. Explained variances for the mathematics subsample is in the numbers above and for the German subsample in the numbers below; R^2 = total percentage of explained variance by intelligence and CPS; ΔR^2 = percentage of variance additionally explained by the particular residual factor beyond the other predictor.
$^*p < .05$.

3.1.4 Discussion

Study 1 revealed three main findings: Fist, the correlation between the good *g* and CPS was (very) high (r = .76/.69). These correlations exceeded the formerly reported coefficients between a ”weak *g*” (only figural reasoning) and CPS in other samples (see Stadler et al., 2015). This result fits well in line with the assumption that the correlation between intelligence and CPS seems to depend on the broadness of the intelligence assessment. For further clarification of this presumption, the analyses of study 1 were repeated with a weakened intelligence operationalization in the sense of figural reasoning. A reduction of the intelligence factor indication to only two figural reasoning subtests caused the correlation coefficients to drop to r = .60/.60. Correlations of this magnitude range within the area of other ”weak *g*”-CPS-correlations as reported in the meta-analysis by Stadler et al. (2015; also see Table 2.1). Second, the coefficients between CPS and the criteria of educational success dropped considerably after controlling for the common variance with intelligence. As a consequence, intelligence prevailed as the superior predictor compared to CPS. Third, CPS only showed a unique effect on the mathematics competence test, whereas intelligence showed substantial increments on all four criteria. Again weakening the broadness of intelligence revealed, as expected, a more powerful CPS as its increment additionally predicted the mathematics grades beyond the ”weak *g*”. Because higher relations between CPS and intelligence appear to go hand in hand with lower CPS-increments, it seems as the high relations between CPS and the criteria of educational success are based on the massive overlap between CPS and broadly assessed intelligence.

Interestingly, CPS exhibited (comparably to intelligence) a differential relation to the school subjects mathematics and German, indicating a higher importance of CPS in the mathematics than in the verbal domain (see Kretzschmar et al., 2014; Schweizer, Wüstenberg, & Greiff, 2013; also see section 4.2.1). In fact, the demands of CPS resemble especially the demands of mathematics (OECD, 2014) and the natural science subjects (Kind, 2013; Klahr & Dunbar, 1988). Strategic behaviors for identifying and verifying basic relations between variables, such as applying VOTAT (Tschirgi, 1980), are essential for mathematical and scientific reasoning and are also fundamental in the CPS knowledge acquisition phase. Thus, a thorough examination of the processes while solving CPS tasks, as well as their relations to intelligence and educational success in different domains is a promising field for further research.

3.2 Study 2: Lotz, Scherer, Greiff, & Sparfeldt (2017)

Lotz, C., Scherer R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence, 64,* 98–112.

In study 2, we investigated the application and adaption of two strategic behaviors across a set of CPS tasks with different demands. Moreover, we examined how intelligence facilitated a more effective strategy use.

3.2.1 Theoretical Background and Hypotheses

As mentioned before in section 2.2.4, it is a major advantage of the computer-based assessment of CPS that detailed information about the problem-solving process are automatically documented. The potential of analyzing these log-files was repeatedly praised but, nevertheless, seldom implemented. Past research on CPS process measures either examined merely program-specific behaviors (Güss et al., 2015; Strohschneider & Güss, 1999) or analyzed more generalized behaviors only for one single task (Greiff, Wüstenberg, & Avvisati, 2015) or averaged across all tasks (Greiff et al., 2016; Wüstenberg, Stadler et al., 2014). Thus, it is still an unanswered research question how problem-solvers apply and adapt universal and domain-general strategic behaviors across a set of several CPS tasks, how they react to changes in the tasks type, and how intelligence is connected with a more effective application and adaption of the strategic behaviors.

Specifically, for solving typical CPS tasks within the MCS approach (Greiff et al., 2012), students have to identify different effect types that pose different demands upon them (Hundertmark, Holt, Fischer, Said, & Fischer, 2015) and, thus, different exploration strategies are optimal. To identify non-dynamic effects, the VOTAT strategy (Tschirgi, 1980) is most effective because it singles out the effects of each problem element. To discover dynamic effects, the NOTAT strategy is optimal because the problem solver could observe without interference how the dynamic system is developing by itself. For being successful in solving CPS tasks, it is crucial to apply VOTAT and NOTAT (Greiff et al., 2016; Kröner et al., 2005) and to increase the frequency of effective strategies while working on a CPS test (Güss et al., 2015). Applying optimal strategic behaviors and adapting them when faced with changes in the task types are reasonable actions and, thus, should be related to intelligence.

In study 2, the first five of the nine MicroDYN tasks contained only non-dynamic effects and could be explored most effectively by applying VOTAT. After the fifth task, a change in the task type occurred and dynamic effects, which are most effectively explored by applying NOTAT, might be present next to the non-dynamic effects. Thus, it was our first goal of study 2 to separately inspected the courses of VOTAT and NOTAT across the task set. Next, we examined the role of intelligence for the application and adaption of the two strategic behaviors across a task set with changing demands.

Hypothesis 1. Examining the course of the relative VOTAT frequencies, we expected a progressive increase across the first five tasks (Hypothesis 1a), a significant drop after the introduction of the dynamic effects (Hypothesis 1b), and an increase across the remaining tasks (Hypothesis 1c). Moreover, we assumed substantial and positive relations between intelligence and the application levels and adaption gradients (Hypothesis 1d).

Hypothesis 2. Examining the course of the relative NOTAT frequencies, we expected no increase across the first five tasks (Hypothesis 2a), a significant rise after the introduction of the dynamic effects (Hypothesis 2b), and an increase across the remaining tasks (Hypothesis 2c). Regarding the relations with intelligence (Hypothesis 2d), we assumed – across the first five tasks – a substantial and negative relation with the application level, but we had no clear expectations regarding the adaption gradient. Referring to the last four tasks, we expected positive relations between intelligence and the application level and adaption gradient.

3.2.2 Methods

Participants and Procedure. Study 2 was based on the sample of study 1[2].

Instruments. Intelligence and MicroDYN were assessed as described in study 1.

Analyses. Hypotheses were examined by latent growth curve models (LGCM; Bollen & Curran, 2006). The change in the task type from non-dynamic to potentially dynamic tasks could be regarded as a sharp change that caused a discontinuous trajectory across the task set. Therefore, a discontinuous LGCM design (Hancock, Harring, & Lawrence, 2013; also referred to as piece-wise LGCM, Diallo & Morin, 2015) with two correlated intercept and slope factors was well-suited to answer the research questions (also see section 4.3.3). Accordingly, the model concerning the relative VOTAT frequencies consisted of two LGCM parts which corresponded to the first five MicroDYN tasks that contained only non-dynamic effects and to the last four tasks in which dynamic effects might have occurred.

[2] $N = 495$; the different sample size resulted from one student who did not work on the CPS test, but on all other variables analyzed in study 1 (Lotz et al., 2016).

Across the first five tasks, a linear increase with a reference point on the fifth task was assumed. Across the last four tasks, we assumed a linear increase with a reference point on the sixth task. For the proposed model, we statistically tested whether the hypothesized discontinuous (vs. continuous) growth curve and the assumed linear (vs. quadratic and vs. no-change) slope specifications represented the data adequately. After determining which of these specifications revealed the superior fit, the model was augmented by a latent intelligence factor.

Referring to Hypothesis (1a) and (1c), the non-standardized means and variances of the slope factors were examined to describe the course across the task set. Responding to Hypothesis (1b), we statistically tested whether the relative VOTAT frequency of the fifth task (right before the task type change) was different from the frequency of the sixth task (right after the task type change) by conducting a Satorra-Bentler corrected χ^2-difference-test. To investigate Hypothesis (1d), correlations of the intelligence factor with the two slope and intercept factors were inspected.

Analyses concerning the NOTAT Hypotheses (2a-d) were conducted analogously.

3.2.3 Results

The descriptive courses of the relative VOTAT and NOTAT frequencies are displayed in Fig. 3.4.

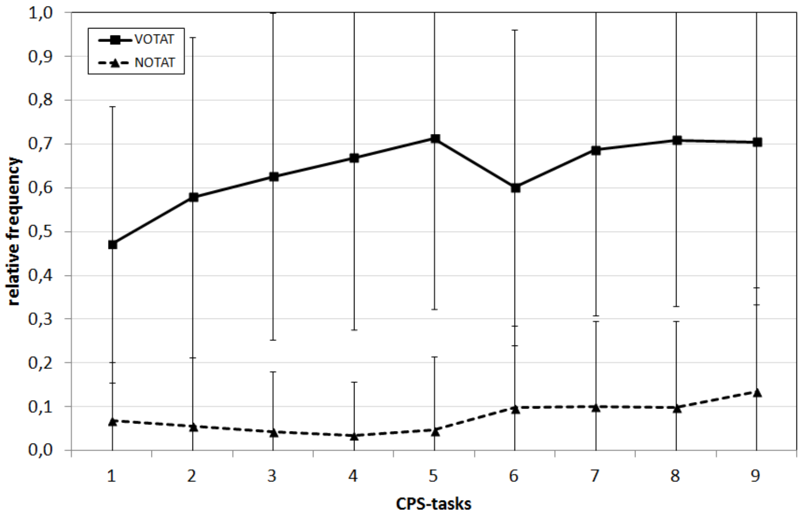


Figure 3.4: Courses of the relative VOTAT and NOTAT frequencies as; manifest means and standard deviations across the nine CPS tasks. Change in task type occurred after task 5.

Results for VOTAT. Among the proposed alternative models of section 3.2.2, the discontinuous LGCM with entirely linear slopes fitted the data most adequately. It was revealed that the relative VOTAT frequency increased significantly across the first five tasks (β = .66; $p < .05$). In accordance with the change in the task type between the fifth and the sixth task, the relative VOTAT frequency dropped significantly ($p < .05$; h = -.15; below the cutoff for small effects; Cohen, 1988) but increased again significantly across the last four tasks (β = .78; $p < .05$). Furthermore, intelligence correlated substantially with the intercept factor before the task type change (r = .48; $p < .05$) and after (r = .40; $p < .05$), indicating that more intelligent students showed higher relative frequency levels of VOTAT throughout the task set. Moreover, a substantial correlation with the slope factor before the task type change (r = .21; $p < .05$) indicated that more intelligent students exhibited a steeper frequency increase across the non-dynamic tasks.

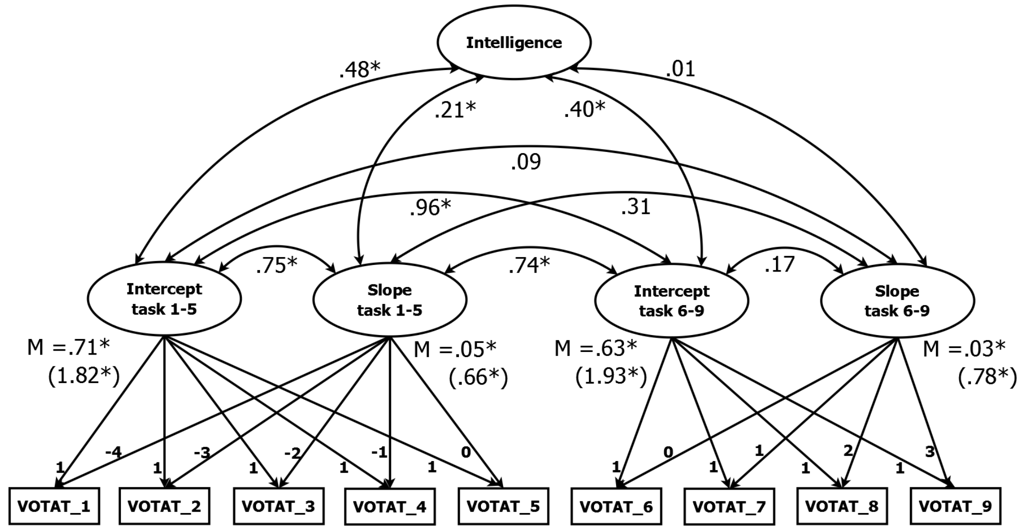


Figure 3.5: Discontinuous latent growth curve model for the relative VOTAT frequency across the nine CPS tasks augmented by intelligence.
Note. Non-standardized (standardized) means are depicted for the intercept and slope factors. Relations among latent variables are shown as correlation coefficients. The measurement model of intelligence is not illustrated.
$^{*}p < .05$.

Results for NOTAT. Among the proposed alternative NOTAT-LGCMs, the discontinuous LGCM with entirely linear slopes also showed the most adequate fit to the data. It was revealed that students showed a slight but significant decrease in their relative NOTAT frequency across the first five tasks (β = -.26; $p < .05$). Corresponding to the change in the task type, the relative NOTAT frequency rose significantly ($p < .05$; h = .24; small effect size; Cohen, 1988) and increased significantly across the last four task as NOTAT became effective (β = .30; $p < .05$). Referring to the relations with intelligence, intelligence correlated significantly negative with the intercept factor before the task type change (r = -.16; $p < .05$), but significantly positive after the task type change (r = .27; $p < .05$). Thus, more intelligent students showed lower levels of NOTAT when this behavior was not effective, but higher levels when this behavior was effective. However, intelligence exhibited no substantial relations to the slope factor before (r = -.14; p = .15) or after the task type change (r = -.05; p = .77).

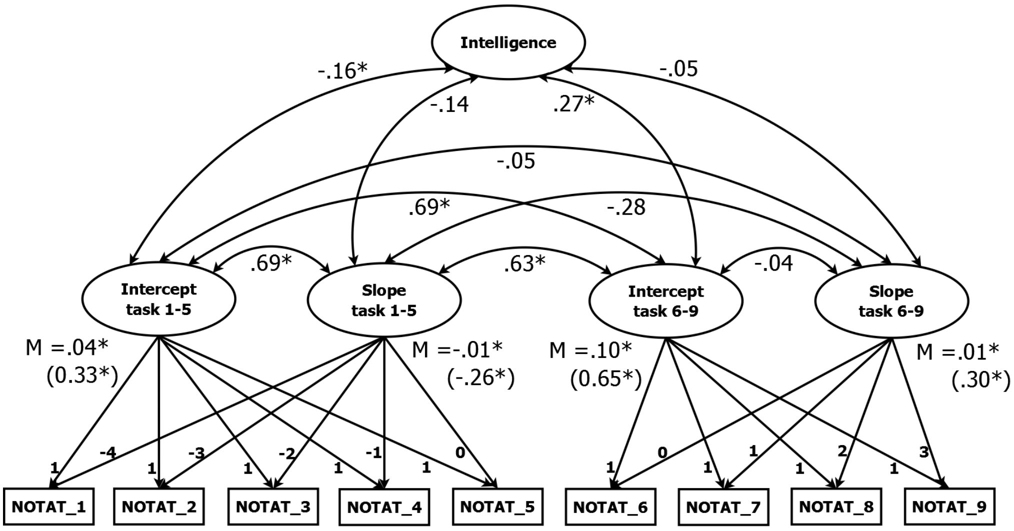
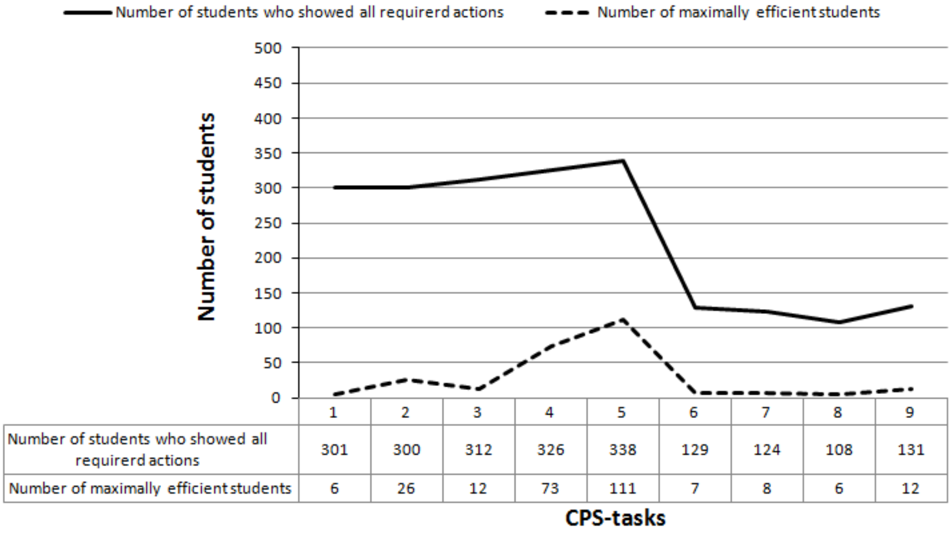


Figure 3.6: Discontinuous latent growth curve model for the relative NOTAT frequency across the nine CPS tasks augmented by intelligence.
Note. Non-standardized (standardized) means are depicted for the intercept and slope factors. Relations among latent variables are shown as correlation coefficients. The measurement model of intelligence is not illustrated.
$^*p < .05$.

3.2.4 Supplementary Analyses

Maximum Efficiency. Next to analyzing the application levels and the adaption gradients of the *effective* VOTAT and NOTAT behaviors, it is also of interest to explore how *efficient* students performed while working on the CPS task set. Therefore, the required exploration steps were identified for each of the nine tasks (see Appendix A). For example, the first MicroDYN task consisted of two input variables and only direct effects. Therefore, VOTAT has to be applied two times (i.e. once for each input variable) to detect all implemented effects. Next, it was determined for each task how many students showed these required exploration steps with all of their exploration steps. As seen in Fig. 3.7, more than half of the students showed the necessary exploration steps (i.e., applied VOTAT for each of the input variables) within all of their exploration step of the first five non-dynamic tasks. Then, the number dropped considerably in correspondence to the task type change: as soon as the tasks potentially contained dynamic effects, only about 25% of the students showed all required exploration steps (i.e., applied NOTAT once in addition to VOTAT for each input variable). Possibly, the majority of the students either did not know NOTAT or were not able to realize that the application of NOTAT was necessary.



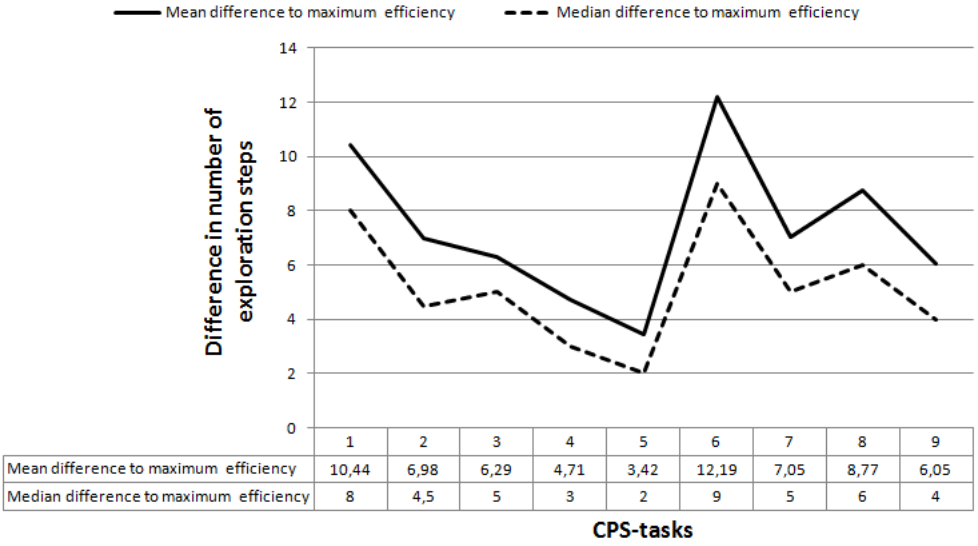
	1	2	3	4	5	6	7	8	9
Number of students who showed all requirerd actions	301	300	312	326	338	129	124	108	131
Number of maximally efficient students	6	26	12	73	111	7	8	6	12

Figure 3.7: Number of students who showed all required exploration steps within all of their exploration steps and number of students who were maximally efficient. Task type change occurred after task 5.

Moreover, the number of students who exactly showed solely the required exploration steps, i.e. were maximally efficient, was computed. Fig. 3.7 shows that across the first five non-dynamic tasks, the number of maximally efficient students increased to about 25% of the sample. Again, the drop in the graph corresponded to the task type change. After dynamic effects potentially occurred, the number of maximally efficient students approached zero and remained on this low level throughout the remaining tasks. Thus, it seems as even those students who were able to apply NOTAT had difficulties to explore the problem space efficiently as the complexity of the problem space increased.

To summarize, across the first five tasks, it seems that the majority of the students were quite proficient problem solvers and that an increasing number of them were able to explore the problem space not only effectively but also efficiently. However, this pattern only held as long as only direct effects were present. As soon as the students were confronted with dynamic effects across the last four tasks, virtually no student was able to exhibit maximally efficient exploration behavior despite some of them were aware of how to apply NOTAT. Possibly, the students struggled with the increasing complexity of the tasks, which was caused by the presence of dynamic effects.

To consider the students' efficient behaviors in more detail, the difference between the number of actually performed exploration steps and the number of maximally efficient exploration steps was computed for those students who performed all required exploration steps within all of their exploration steps. As seen in Fig 3.8 the mean difference as well as the median difference of exploration steps to being maximally efficient decreased progressively across the first five non-dynamic tasks. Thus, students progressively approached towards being more efficient. In accordance with the task type change, the differences rose again considerably from task 5 to task 6 as dynamic effects could have occurred. Across the potentially dynamic tasks, again an overall decrease of the differences could be documented, indicating that students again became more efficient. Interestingly, the differences in task 7 were lower than in task 8. This is due to task 7 being the only task within the potentially dynamic tasks that, in fact, did not contain dynamic effects. Assumingly, students had less difficulties to explore the problem space in the absence of dynamic effects and, thus, were slightly more efficient.



	1	2	3	4	5	6	7	8	9
Mean difference to maximum efficiency	10,44	6,98	6,29	4,71	3,42	12,19	7,05	8,77	6,05
Median difference to maximum efficiency	8	4,5	5	3	2	9	5	6	4

Figure 3.8: Difference between actually performed exploration steps and maximum efficiency computed for only those students who showed all required actions. Task type change after task 5.

To conclude, although the tasks became more and more complex, the students showed a slow but steady progress towards being more efficient across the non-dynamic tasks and – after a considerable relapse towards being less efficient because of the task type change – they continued that trend towards being more efficient also across the dynamic tasks. However, the absolute number of maximally efficient students was low, especially if dynamic effects were present. Assumingly, the occurrence of dynamic effects increased the complexity of the tasks and might have caused more extensive exploration behavior.

As a cautionary note, it has to be kept in mind that the students were not explicitly instructed to show maximally efficient exploration behavior. Thus, it is theoretically conceivable that the students would have shown more efficient exploration behavior if they would have been instructed to try to explore the problem space by only applying the minimal amount of necessary exploration steps.

Strategic Behaviors as Predictors of Educational Success. Past research gave empirical evidence for the importance of VOTAT in educational contexts beyond intelligence (Bryant et al., 2015). Moreover, as the results of study 2 suggested (see section 3.2.3), it is a next crucial step in research on strategic CPS behavior to examine the incremental validity of the exploration behavior for real-life criteria such as educational success beyond intelligence. In comparison to conventional intelligence tests, CPS tests feature an interactive exploration phase. Thus, Lotz et al. (2017) proposed that the specific type of interaction behavior during the exploration phase might represent the assumed unique aspects of CPS tests compared to conventional intelligence tests. As a hypothesis, the strategic exploration behaviors VOTAT and NOTAT might predict the criteria of educational success beyond intelligence.

All following analyses were run separately for the two subsamples as described in study 1 (see section 3.1.2). Concerning the increment of VOTAT, two residual models (one for each subsample) in which intelligence and the VOTAT-residual (increment which is independent from intelligence) were used as predictors of the scholastic competence tests and grades. Thereby, the VOTAT factor was indicated by the nine relative VOTAT frequency items[3]. For examining the NOTAT increment, analogous models were computed in which the NOTAT factor was indicated by the nine relative NOTAT frequency items. Because of directed hypotheses, tests were conducted one-sided.

[3] Unfortunately, the latent factors that represented the application levels and adaption gradients of VOTAT in the LGCM of study 2 could not be used as predictors of educational success because a very high correlation between the intercept factors ($r_{\text{VOTAT Intercepts}}$ = .96) caused massive multicollineraty problems. Comparable multicollineraties were also present in the NOTAT-LGCM.

Referring to VOTAT, the residual model of the mathematics subsample fitted acceptably (χ^2 = 1216.154; *df* = 1167; CFI = .917; TLI = .913; RMSEA = .013) and corresponding results are shown in Fig. 3.9. VOTAT and intelligence correlated substantially (*r* = .39; $p < .05$). Intelligence substantially explained 72.3% (β = .85; $p < .05$) of the mathematics competence test variance, whereas the increment of VOTAT non-significantly added less than one percent (β = .05; *p* = .48), resulting in a total of 72.4% explained variance. Comparably, intelligence substantially explained 17.2% (β = .42; $p < .05$) of the mathematics grade variance, whereas the increment of VOTAT again added less than one percent (β = .06; *p* = .41), resulting in a total of 17.5% explained variance. Concerning the German subsample, the VOTAT-residual model did not reach acceptable fit statistics (χ^2 = 535.940; *df* = 222; CFI = .870; TLI = .852; RMSEA = .075; Schermelleh-Engel, Moosbrugger, & Müller, 2003). Therefore, results are not reported in detail.

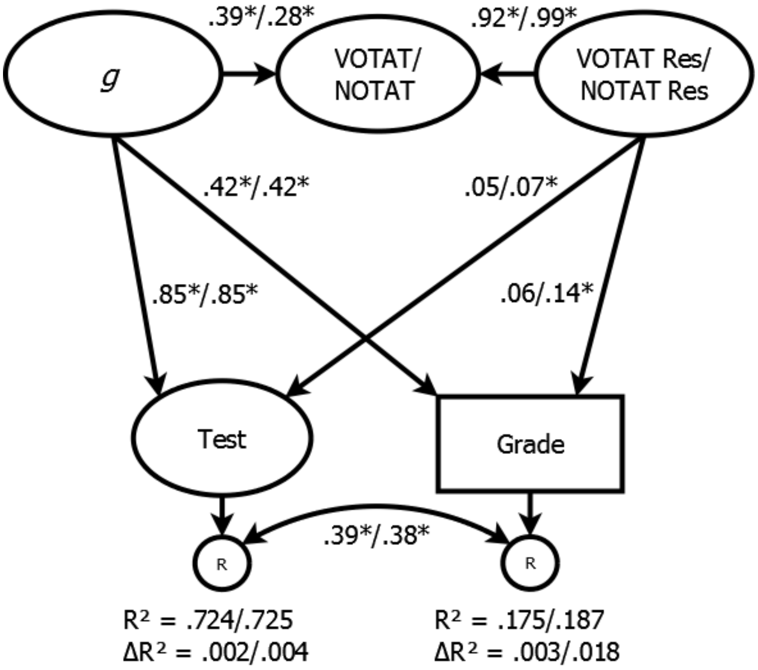


Figure 3.9: Standardized coefficients of the VOTAT increment model/NOTAT increment model in mathematics. Scholastic competence tests and reversely scored grades in mathematics were regressed on intelligence and the increment of VOTAT/NOTAT. R^2 = total percentage of explained variance by intelligence and VOTAT/NOTAT; ΔR^2 = percentage of variance additionally explained by the increments of VOTAT/NOTAT beyond intelligence.
$^*p < .05$.

Referring to NOTAT, comparable residual models were specified. However, the initial model in which the latent NOTAT factor was indicated by nine relative NOTAT frequency items revealed very inadequate fit statistics for the mathematics as well as the German subsample (e.g., CFI $< .800$; TLI $< .800$; RMSEA $> .100$). A subsequently specified model in which the latent NOTAT factor was indicated by those four items that corresponded to the tasks with the potential occurrence of dynamic effects (tasks 6-9) showed partly better fit statistics.

With regard to the mathematics subsample, the NOTAT-residual model fitted acceptably (χ^2 = 978.735; *df* = 937; CFI = .921; TLI = .917; RMSEA = .014) and corresponding results are displayed in Fig. 3.9. NOTAT and intelligence correlated substantially (*r* = .28; $p < .05$). Intelligence substantially explained 72.1% (β = .85; $p < .05$) of the mathematics competence test variance and the increment of NOTAT significantly added 0.4% of variance (β = .07; $p < .05$), resulting in a total of 72.5% explained variance. Comparably, intelligence substantially explained 16.9% (β = .41; $p < .05$) of the mathematics grade variance, whereas the increment of NOTAT significantly added 1.8% (β = .14; $p < .05$), resulting in a total of 18.7% explained variance. Referring to the German subsample, the model still fitted inadequately (χ^2 = 236.146; *df* = 127; CFI = .828; TLI = .793; RMSEA = .059; Schermelleh-Engel et al., 2003). Thus, results are neither reported in detail nor interpreted.

3.2.5 Discussion

Study 2 revealed three main findings: First, students showed higher application rates of a strategic behavior if this behavior was effective, but lower application rates if the strategic behavior was not effective. Second, students progressively adapted their strategic behaviors across the task set and reacted to changes in the task type. Specifically, they showed increasing application gradients if a behavior was effective, but deceasing application gradients if a behavior was not effective. Third, intelligence was substantially related to the application and adaption of strategic behaviors. More intelligent students applied higher levels of VOTAT and NOTAT if these behaviors were effective, but lower levels if the behaviors were not. Additionally, students showed a steeper gradient when adapting their VOTAT behavior across the tasks 1-5.

Concerning the VOTAT results of study 2, it is especially noteworthy that the relative frequency was steadily increasing throughout both task types. This gradual increase is in accordance with the results of Schauble (1996) and Vollmeyer et al. (1996). Students were not simply abandoning ineffective strategies. Instead, they were changing their exploration behavior rather slowly but progressively. Moreover, the relative VOTAT frequency increased in both task types until it reached about 70%, but it never exceeded this level. Thus, one might argue that the increasing trend could have reached an upper boundary. In face of real-life contexts, which confront students with novel and uncertain situations that might be subject to unexpected and randomly occurring changes, such a behavior seems reasonable. Instead of rigidly adhering to one specific strategy, students might have tested a set of different strategic behaviors such as VOTAT, NOTAT, or maybe even HOTAT (hold-one-hing-at-a-time, optimal strategy to detect interaction effects; Tschirgi, 1980). Such a variety of systematic strategical behaviors would lead to more success in real-life situations, as more intelligent people are known to have (e.g., Sternberg, Grigorenko, & Bundy, 2001).

Concerning the NOTAT results of study 2, it is noteworthy that the NOTAT intercept-intelligence-correlations showed a differential pattern of results: across the non-dynamic tasks, the coefficient was negative, but across the potentially dynamic tasks, the coefficient was positive. This pattern indicated that higher intelligence might have enabled students to better identify which behaviors were effective to explore the unknown problem space and to flexibly adapt their application levels to changing task demands.

In the supplementary analyses, the maximally efficient behaviors for each task were analyzed. Similarly to the examination of effectiveness, the students showed a slow but steady progress towards being more efficient, although the tasks were becoming more and more complex within the particular tasks types. Probably, a process that might be comparable to learning enabled the students to progress from applying many and probably non-effective or redundant exploration steps to applying the necessary effective behaviors.

Nevertheless, these results should be interpreted with caution because students were solely instructed to explore the problem space, but not to explore the problem space most efficiently. The low number of maximum efficient students, especially across the potentially dynamic tasks, might stem from the anticipation of another task type change that might not have been introduced. Thus, in the context of these MicroDYN tasks, students might have got the impression that it is rather effective to apply a wider range of strategic behaviors which prevented them from being maximum efficient. Another reason for the low number of maximum efficient students might be that the students frequently conducted double experiments. By repeating the same variable manipulation several times instead of referring back to prior manipulations, students might have applied a very convenient way to unload their working memory capacity. In turn, this entirely reasonable behavior might have lead to a higher number of variable manipulations as would have been necessary to completely explore the problem space.

Additionally, the supplementary analyses also concerned the prediction of educational success by the strategic behaviors beyond intelligence. Unfortunately, the residual models of VOTAT and NOTAT revealed only partly acceptable fitting models. Referring to those models of the mathematics subsample that exhibited sufficient fit statistics, the analyses showed that VOTAT had no substantial added value beyond intelligence for scholastic competence tests and grades in mathematics. However, NOTAT exhibited a small but substantial effect on both criteria. Thereby, it is especially noteworthy, that the latent NOTAT factor was only indicated by those four items that corresponded to the tasks with potentially dynamic effects. Possibly, the ability of dealing with dynamic effects is of higher importance for real-life outcomes than previously assumed. The results of this supplementary analysis fit well in the line of results that indicated NOTAT's importance (see Greiff et al., 2016; Schoppek & Fischer, 2017). Moreover, they augmented the current state of research with first results about NOTAT's significance for real-life outcomes.

3.3 Study 3: Lotz, Schneider, & Sparfeldt (2018)

Lotz, C., Schneider, R., & Sparfeldt, J. R. (2018). Are intelligence and motivation differentially relevant for scholastic competence tests and grades in mathematics? *Learning and Individual Differences, 65*, 30-40.

In study 3, we statistically examined the differential relevance of intelligence, self-concept, and interest for scholastic competence tests and grades in mathematics. Furthermore, we focused on the unique effects that each predictor exhibited on educational success beyond the other predictors.

3.3.1 Theoretical Background and Hypotheses

Although educational success is largely determined by intelligence (e.g., Jensen, 1998a), students are not always performing as good as they are actually capable of. In study 1 and in the supplementary analyses of study 2, performance measures as well as process measures of CPS revealed rather low amounts of additionally explained variance beyond intelligence for predicting educational success. Thus, widening the search space and taking a closer look at non-cognitive variables such as motivation might be a more fruitful approach.

In past research, motivational variables such as academic self-concept and interest were frequently shown to account for substantial amounts of variance beyond intelligence (e.g., Kriegbaum et al., 2015; Spinath et al., 2006; Steinmayr & Spinath, 2009). However, it seems that motivational variables as well as cognitive variables show a differential relevance, depending on which type of achievement indicator (scholastic competence tests or grades) was used: Whereas motivation seems to be more relevant than intelligence for grades, intelligence seems to be more relevant than motivation for scholastic competence tests (Helmke, 1992;

Jansen et al., 2016; Steinmayr & Meißner, 2013). Nevertheless, most of these prior studies, claiming the described differential prediction pattern, reported the differences between the particular regression coefficient only numerically, but did not use inferential statistical tests. Up to now, only the study by Steinmayr and Meißner (2013) statistically compared the coefficients of intelligence and mathematics self-concept for mathematics competence tests and mathematics grades. It was revealed that intelligence showed higher coefficients than self-concept on mathematics competence tests, but intelligence and self-concept did not differ significantly in their coefficients on mathematics grades although the numerical pattern was almost perfectly inverted. Concerning interest, the differential prediction pattern was not yet statistically examined. Moreover, previous studies that gave numerical evidence for the differential prediction pattern did not control for the large amounts of shared variance between the predictors intelligence, self-concept, and interest. Thus, the examination of the unique effects of each predictor beyond the others would further deepen the understanding of the interplay of cognitive and motivational variables as predictors of different indicators of educational success.

In the light of these points, we examined the following hypotheses for the school subject mathematics:

Hypothesis 1. Examining the differential relevance of solely intelligence and self-concept, we expected that intelligence was more relevant than self-concept for the mathematics competence test, whereas self-concept was more relevant than intelligence for grades.

Hypothesis 2. Examining the differential relevance of solely intelligence and interest, we expected that intelligence was more relevant than interest for the mathematics competence test, whereas interest was more relevant than intelligence for grades.

Hypothesis 3. Examining the differential relevance of intelligence, self-concept, and interest in concert, we assumed that the differential prediction pattern of intelligence and self-concept remained as expected in Hypothesis (1), whereas the differential relevance of interest might vanish.

Hypothesis 4. Again considering all three predictors in concert, we expected that the increment of intelligence (Hypothesis 4a) and self-concept (Hypothesis 4b) was substantial. Whether the increment of interest (Hypothesis 4c) showed unique effects was an open research question.

3.3.2 Methods

Participants and Procedure. Study 3 was based on the mathematics subsample of study 1 (N = 245; n = 120 females; mean age = 16.39, SD = 0.95).

Instruments. Intelligence, the mathematics competence test, and the mathematics grade were assessed as described in study 1 (see section 3.1.2). Mathematics self-concept was assessed by the Differential Self-Concept Grid (DISC-Grid; Rost et al., 2007) and mathematics interest was measured with four items (Sparfeldt, Rost, & Schilling, 2004).

Analyses. For statistically comparing path coefficients, it is a mayor obstacle that all variables were assessed with different metrics. Thus, only the standardized path coefficients that are not affected by the metric of their scales could be compared meaningfully. However, no common statistics software (such as M*plus*) has a built-in function to compare standardized coefficients. Equality constraints to test the difference between two or more coefficients by a likelihood ratio test could only be imposed on the non-standardized coefficients. However, comparing non-standardized coefficients results in the mentioned interpretation problems due to different metrics. The two-stage method proposed by Kwan and Chan (2011) overcomes this obstacle. At stage 1, the original model's standardized paths (that are meaningfully comparable but not testable against each other) were transformed into the non-standardized paths of the transformed model which is covariance-equivalent to the original model. At stage 2, equality constraints were posed upon the non-standardized path coefficients of the transformed model (that are equal to the standardized path coefficients of the original model) to statistically test their difference via Wald tests.

Accordingly, we employed this method to analyze Hypotheses (1) to (3). Referring to Hypothesis (1), a latent model with intelligence and mathematics self-concept as predictors and the mathematics competence test and grades as criteria was established. After transforming this model, equality constraints were subsequently imposed on the paths of interest to test the assumed differential prediction pattern. Analyses concerning Hypothesis (2) were conducted analogously with a model that contained intelligence and mathematics interest as predictors. Regarding Hypothesis (3), intelligence, mathematics self-concept, and mathematics interest predicted the criteria in concert. Again, the original model was transformed and equality constraints were subsequently imposed on the parameters of interest to test the assumed differential relevance.

For examining the increments of each predictor, we employed the Cholesky factoring approach (de Jong, 1999; Loehlin, 1996). By orthogonally decomposing the explained variances, a hierarchical regression analysis could be performed in only one model without altering the model fit or affecting the measurement part of the model. To examine the intelligence increment (Hypothesis 4a), self-concept was assigned first priority, second priority was assigned to interest and last priority was assigned to intelligence. Thus, the intelligence Cholesky factor represented the increment of intelligence after the shared variance with self-concept and interest was partialled out. Comparably, to examine the self-concept increment (Hypothesis 4b), we assigned intelligence first, interest second, and self-concept last priority. To examine the interest increment (Hypothesis 4c), we assigned intelligence first, self-concept second, and interest last priority. An inspection of the regression coefficients of the particular Cholesky factors would indicate whether the increment is significant. Furthermore, by squaring the regression coefficients, a straightforward inspection of explained variances is allowed.

3.3.3 Results

As expected in Hypothesis (1), intelligence was significantly more relevant than mathematics self-concept for the mathematics competence test (.74 vs. .23; $p < .05$), whereas self-concept was significantly more relevant than intelligence for grades (.57 vs. .16; $p < .05$), indicating the assumed differential prediction pattern (see left side of Fig. 3.10). Moreover, intelligence predicted the mathematics competence test significantly higher than the mathematics grade (.74 vs. .16; $p < .05$), whereas self-concept predicted the mathematics grade significantly higher than the mathematics competence test (.57 vs. .23; $p < .05$).

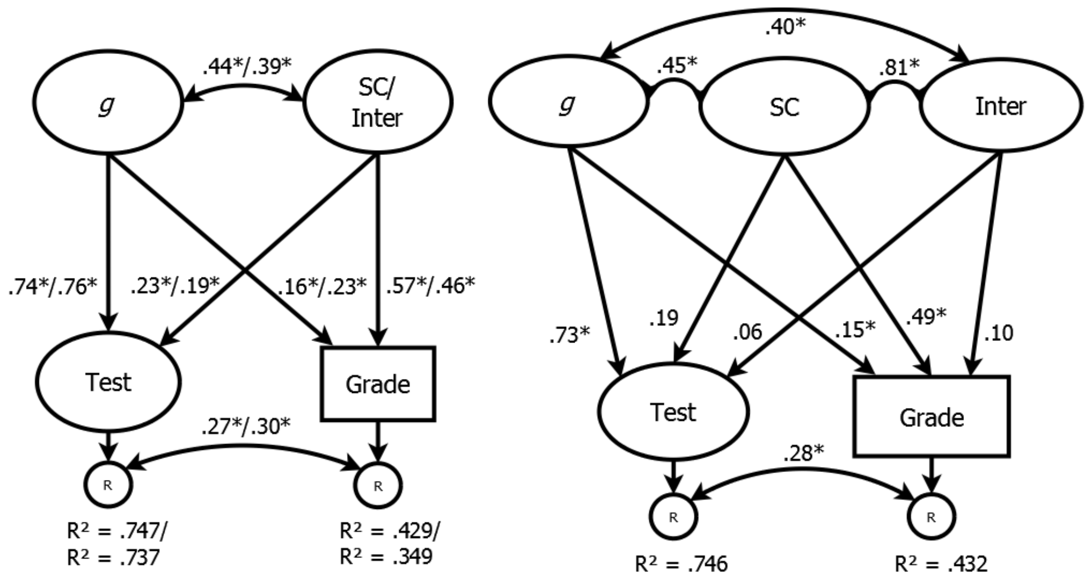


Figure 3.10: Structural model with standardized path coefficients of intelligence and mathematics self-concept/interest (left model) and structural model with standardized path coefficients of intelligence, mathematics self-concept, and mathematics interest (right model). Measurement models were not depicted. R^2 = total percentage of explained variance.
$^*p < .05$.

Comparably, also for mathematics interest and intelligence (Hypothesis 2; also see left side of Fig. 3.10) the differential prediction pattern was revealed by significantly higher regression coefficients of intelligence than interest on the mathematics competence test (.76 vs. .19; $p < .05$), but significantly higher regression coefficients of interest than intelligence on mathematics grades (.46 vs. .23; $p < .05$). Furthermore, intelligence revealed to be significantly more important for the mathematics competence test compared to grades (.76 vs. .23; $p < .05$), whereas interest revealed to be significantly more important for grades compared to the mathematics competence test (.46 vs. .19; $p < .05$).

When all three predictors were regarded in concert (Hypothesis 3; see right side of Fig. 3.10), the differential prediction pattern also emerged. Results indicated that intelligence exhibited significantly higher regression coefficients than self-concept (.73 vs. .19; $p < .05$) and interest (.73 vs. .06; $p < .05$) on the mathematics competence test, whereas self-concept and interest did not differ significantly from each other (.19 vs. .06; ; $p = .57$). In turn, self-concept revealed significantly higher coefficients than intelligence (.49 vs. .15; $p < .05$) and interest (.49 vs. .10; $p < .05$) on grades, whereas intelligence and interest did not differ significantly from each other (.15 vs. .10; $p = .63$). Moreover, intelligence again predicted the mathematics competence test significantly higher than grades (.73 vs. .15; $p < .05$), but self-concept predicted the mathematics grade significantly higher than the mathematics competence test (.49 vs. .19; $p < .05$). Interest was comparably non-predictive for both criteria (.06 vs. .10; $p = .74$).

Regarding the unique aspects of each predictor, from a total of 74.6% explained variance of the mathematics competence test, the intelligence increment (Hypothesis 4a; see Fig. 3.11) significantly explained 42.3% ($\beta = .65$; $p < .05$) additional variance beyond self-concept and interest. From a total of 43.2% explained variance of the mathematics grade, the intelligence increment significantly added 1.7% ($\beta = .13$; $p < .05$). Referring to the self-concept increment (Hypothesis 4b; see Fig. 3.12), it added no significant amount of variance to the mathematics competence test beyond intelligence and interest (1.8%; $\beta = .11$; $p = .10$), but it significantly contributed 8.1% ($\beta = .28$; $p < .05$) to the mathematics grade. In contrast, the unique aspects of interest (Hypothesis 4c; see Fig. 3.13) non-significantly added less than 1% to the prediction of the mathematics competence test ($\beta = .04$; $p = .63$) as well as grades ($\beta = .06$; $p = .23$) beyond intelligence and self-concept.

3.3.4 Supplementary Analyses

Unlike studies 1 and 2, study 3 did not utilize the full sample to answer the research questions, but only the mathematics subsample. The reason for this proceeding was that the measures of motivation and the measures of achievement did not align to each other in the same way for both subsamples. In the mathematics subsample,

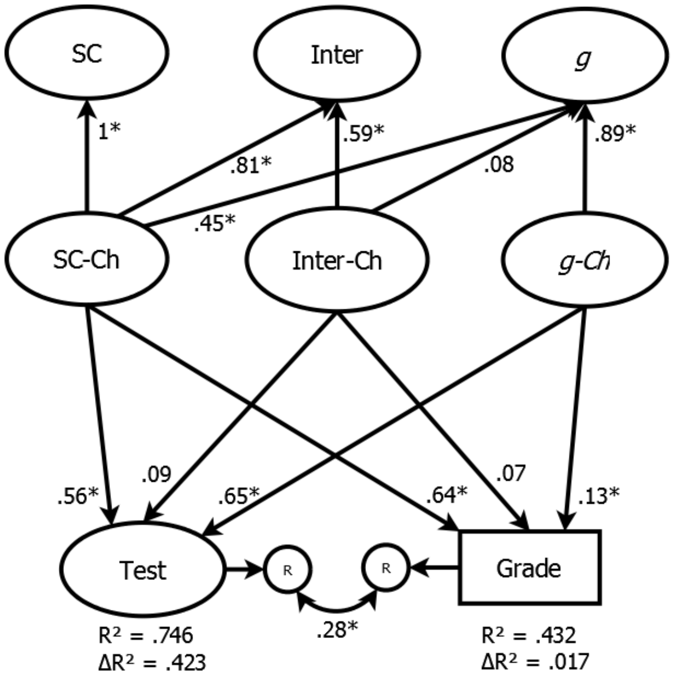


Figure 3.11: Cholesky factoring model of the intelligence increment of the mathematics subsample. Measurement models were not depicted. R^2 = total percentage of explained variance; ΔR^2 = percentage of variance additionally explained by the intelligence increment beyond self-concept and interest.
$^*p < .05$.

the measures to assess mathematics motivation and mathematics achievement both referred to the school subject mathematics as a whole. However, in the German subsample, only the self-concept and interest measures referred to the school subject German as a whole, but the scholastic competence test in German only tapped the facet of reading comprehension. Predicting a criterion that is not as broad as the predictor might cause an asymmetry that could distort the result pattern (Wittmann, 1988). Thus, analyses for the German subsample were not included in study 3. Nevertheless, for reasons of completeness, the following section gives a short overview about the analyses of the German subsample.

Methods. The sample under investigation corresponded to the German subsample of study 1 (see section 3.1.2; N = 251 students; n = 108 females; mean age = 16.40, SD = 0.93). With reference to the instruments, intelligence, the reading comprehension test, and the German grade were assessed as described in study 1 (see section 3.1.2). The German self-concept was assessed by the Differential Self-Concept Grid (DISC-Grid; Rost et al., 2007; Cronbach's α = .89) and German

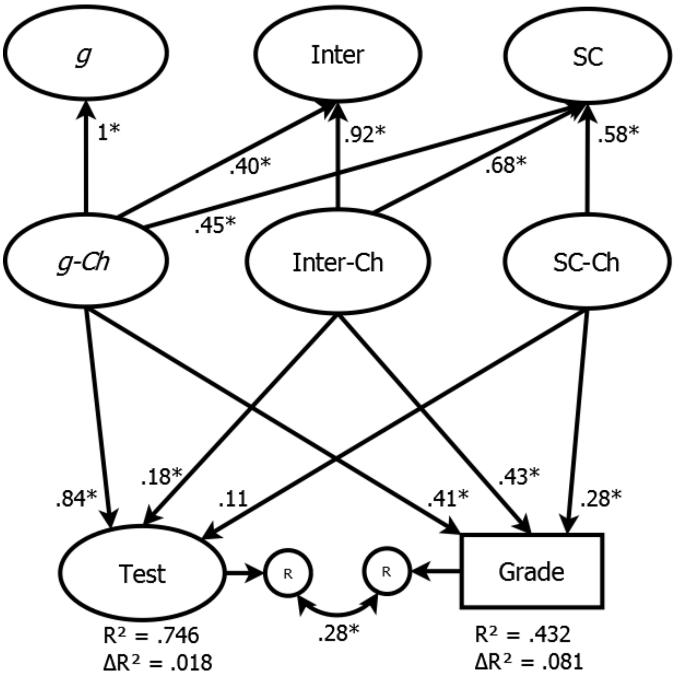


Figure 3.12: Cholesky factoring model of the mathematics self-concept increment of the mathematics subsample. Measurement models were not depicted. R^2 = total percentage of explained variance; ΔR^2 = percentage of variance additionally explained by the self-concept increment beyond intelligence and interest.
$^*p < .05$.

interest was measured by four items (Sparfeldt et al., 2004; Cronbach's α = .84). Regarding the analyses, all models were conducted analogously to study 3 with the German self-concept and German interest items indicating the predictors as well as the reading comprehension items and the German grade as indicators of the criteria.

Results. The latent model that solely contained intelligence and German self-concept did not reach acceptable fit indices (χ^2 = 900.624; *df* = 835; CFI = .867; TLI = .856; RMSEA = .018). Comparably, the latent model that solely contained intelligence and German interest also failed to fit at least acceptably (χ^2 = 715.998; *df* = 677; CFI = .852; TLI = .838; RMSEA = .015). The model that regarded all three predictors in concert also revealed fit statistics that did not reach an acceptable magnitude (χ^2 = 1074.364; *df* = 1005; CFI = .876; TLI = .866; RMSEA = .017). Accordingly, the Cholesky factoring models showed the same insufficient fit statistics as these models were covariance-equivalent to the model with the three predictors.

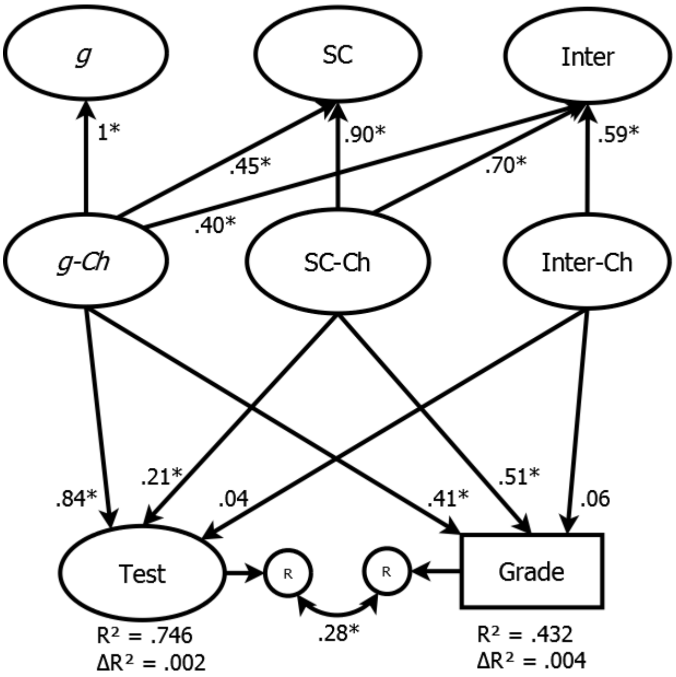


Figure 3.13: Cholesky factoring model of the mathematics interest increment of the mathematics subsample. Measurement models were not depicted. R^2 = total percentage of explained variance; ΔR^2 = percentage of variance additionally explained by the interest increment beyond intelligence and self-concept.
$^*p < .05$.

As a consequence of these overall insufficient fit indices (Schermelleh-Engel et al., 2003), results of the models were not reported in detail because a meaningful testing and interpretation of the differential prediction pattern and the unique effects of the predictors was not warranted for the models of the German subsample.

3.3.5 Discussion

Study 3 revealed four main findings: First, when solely regarding intelligence and mathematics self-concept, intelligence was the significantly stronger predictor than self-concept for the mathematics competence test, whereas self-concept was the significantly stronger predictor for grades. Second, when solely intelligence and mathematics interest were considered, a similar pattern of results was obtained. Third, when analyzing all three predictors in concert, intelligence prevailed as the strongest of the considered predictors for the mathematics competence test, whereas self-concept remained as the strongest for mathematics grades. The interest coefficients dropped to around zero and the differential importance of interest vanished completely. Fourth, concerning the specific increments, intelligence showed unique effects beyond both motivational variables on both criteria, thereby the increments on the mathematics competence test were numerically larger than those on grades. In turn, the self-concept increment revealed substantial unique effects beyond intelligence and interest on the mathematics grade, but not on the mathematics competence test. In contrast, interest did not show a substantial increment beyond the other two predictors.

It is especially noteworthy, that this study was the first to statistically compare the regression coefficients of intelligence, self-concept, and interest on scholastic competence tests and grade in one sample and by applying advanced statistical methods. The finding that intelligence and self-concept are differentially relevant was in accordance with the numerically reported pattern by Helmke (1992), but, however, partly contradicted the pattern of Steinmayr and Meißner (2013). Using less advanced methods for conducting the inferential statistical path comparisons, they showed that the differential pattern was significant for mathematics competence tests but non-significant for grades, although the numerical patterns were almost perfectly inverted. One methodological reason might be the different analysis approaches: whereas Steinmayr and Meißner (2013) provided no information about how they dealt with the different metrics of the compared paths, study 3 used the two-step approach to handle this obstacle. Another reason could be located in the sample characteristics: whereas Steinmayr and Meißner (2013) assessed thirteen-year-old students, we examined sixteen-year-old students. Because the relation between intelligence and grades seems to decrease with age (Jensen, 1998a), self-concepts become more differentiated and stable with age which leads to increasing relations with grades (Skaalvik & Valås, 1999). Thus, the impact of motivation on achievement-related behavior has possibly not been fully developed in the younger high school student sample of Steinmayr and Meißner (2013), whereas self-concept might have steered the achievement-related behavior considerably stronger. As a consequence, the significantly higher relations between self-concept and grades than between intelligence and grades in study 3 might have been a result of the students' different age ranges.

Chapter 4

General Discussion

In this dissertation project, it was aimed to shed light on the incremental validity that cognitive, strategic-behavioral, and motivational variables show beyond broadly operationalized intelligence when predicting educational success in the school subjects mathematics and German assessed with scholastic competence tests and grades as achievement indicators. The following chapter summarizes the main findings of this dissertation project, reflects on the results' implications for research in educational contexts, and critically acknowledges the key features of the three studies and the design of this dissertation project.

4.1 Summary of Main Findings

Across the three studies of this dissertation project, the extraordinary role of intelligence as one of the most important psychological determinants of educational success was acknowledged. Nevertheless, there is still room left for other constructs to uniquely contribute to the prediction of educational success beyond intelligence. Accordingly, the three studies revealed that the majority of the considered predictors additionally explained mostly small but substantial amounts of unique variance in some of the assessed criteria of educational success beyond broadly operationalized intelligence. As an exception, academic self-concept even exhibited considerable unique effects beyond intelligence for grades. Furthermore, the results of the studies 1 and 2 conduced to the current debate about the relation between intelligence and CPS: Regarding performance measures of CPS, additional empirical evidence showed that the relation between intelligence and CPS depends on the broadness of the intelligence assessment. Referring to process measures of CPS, it was revealed that intelligence facilitates effective strategic exploration behavior.

Specifically, in study 1, we regarded the incremental validity of the cognitive construct of CPS, by focusing on its performance measures. The relation between intelligence and CPS performance measures was the closer, the broader intelligence was assessed. Moreover, the formerly substantial correlations between CPS and scholastic competence tests and grades dropped considerably after controlling for the huge amounts of shared variance between intelligence and CPS. Subsequent path comparisons revealed that intelligence was the superior predictor for all four achievement measures. As an augmentation of all prior studies, it was revealed that CPS only showed a substantial increment on the mathematics competence test, but not on the reading comprehension test nor on the grades in mathematics or German beyond broad measures of intelligence. In contrast, the comprehensive intelligence measure exhibited substantial unique effects beyond CPS on all four achievement indicators. Thus, broader intelligence seemed to have more unique variance in common with educational success than CPS. It is highly probable that the huge overlap between intelligence and CPS was responsible for this pattern of results, by lowering the probability of a substantial CPS increment.

In study 2, we focused on the strategic-behavioral process measures of CPS. Concerning the course of VOTAT and NOTAT across the task set with changing demands, students showed higher application frequencies of effective strategic behaviors and progressively adapted their behaviors. Moreover, intelligence was positively related to higher application frequencies and partly to steeper adaptions gradient of optimal behaviors, indicating that intelligence manifested itself in effective strategic behaviors. Supplementary analyses revealed no substantial contribution of VOTAT to the prediction of educational success beyond intelligence. Contrarily, NOTAT incrementally added small but substantial amounts of variance to the prediction of the mathematics competence test and mathematics grades. Thereby, the rare prior research on NOTAT's importance was augmented by empirical evidence about its significance for real-life outcomes.

In study 3, attention was turned to examining the incremental validity of motivational variables for scholastic competence tests and grades in mathematics. When considering intelligence in combination with solely self-concept or solely interest, all considered variables were substantial predictors of educational success in mathematics and exhibited a differential relevance pattern for scholastic competence tests and grades. As assumed, intelligence was the stronger predictor for the mathematics competence test, whereas the motivational variables were the stronger predictors for mathematics grades. However, when controlling for the shared variance between both motivational variables, only intelligence and self-concept remained as substantial predictors and continued showing the differential relevance pattern. The coefficients of interest dropped to zero and, accordingly, its differential relevance disappeared. Examining the unique aspects of the predictors, intelligence exhibited substantial increments on both criteria and self-concept on the mathematics grades. Interest showed no unique effects beyond intelligence and self-concept. This pattern indicated that the predictive value of interest was merely based on its overlap with self-concept. Contrarily, the outstanding importance of self-concept for grades next to intelligence was fostered.

4.2 Implications

The following section focuses on the implications that could be derived from the results of this dissertation project. Thereby, the implications for research on CPS and motivation in educational contexts as well as the implications for assessing educational success with different achievement indicators are highlighted.

4.2.1 Implications for Research on Complex Problem Solving in Educational Contexts

In study 1, high correlations between broadly assessed intelligence and CPS performance measures were revealed. Correlations of this magnitude are typically obtained for test-retest-estimates or range in the area in which different intelligence tests correlate among each other (Johnson et al., 2008; Valerius & Sparfeldt, 2014). Thus, CPS shared, as expected, very high amounts of variance with intelligence. Nevertheless, CPS showed a rather small but substantial increment on the mathematics competence test. Therefore, it could be concluded that CPS exhibited at least some unique aspects when predicting educational success beyond broadly operationalized intelligence.

In accordance with prior studies, CPS exhibited larger unique effects beyond narrow indications of intelligence, but – for the most part – not beyond broad indications (also see Kretzschmar et al., 2016). Because the incremental effects of CPS beyond broad measures of intelligence fell short of expectations, the question might arise whether educational success is the right criterion to examine the theoretically assumed unique aspects of CPS. Possibly, other 21st century skills could be more appropriate criteria as they might better reflect the higher-order thinking skills that CPS claims to tap. For example, skills like critical thinking, programming, resource management, or teamwork are frequently mentioned as important 21st century skills (Pellegrino & Hilton, 2012). As a speculation, the assumed unique aspects of CPS might tap these higher-order thinking skills better than it taps educational success. However, comprehensive studies that examine the incremental validity of CPS for this kind of real-life criteria are still missing, but they could be a fruitful direction for future research on CPS.

Concerning the question, whether CPS is a separate construct independent of intelligence, still no final conclusion about the exact location within the nomological network of intelligence could be provided by the results of study 1. Possibly, alternative factor model approaches such as nested-factor modeling might provide better insights in the relation between intelligence and CPS (Kretzschmar et al., 2016; Sonnleitner et al., 2013). For example, in the study of Kretzschmar et al. (2016), it was revealed that models with a specific CPS factor in addition to the established intelligence constructs of reasoning, mental speed, memory, creativity, and general knowledge fitted significantly better than models without a specific CPS factor. This pointed towards the distinctness of intelligence and CPS. Nevertheless, future studies should complement such analyses with omegaHS to determine the effect size of the unique latent variable strength of the specific CPS factor (Gignac & Kretzschmar, 2017).

Another point of view implies that the question about the independence of CPS could not be finally answered until a precise process model of CPS activities is established (Funke & Frensch, 2007). Thus, as no such process models exist, it remains unclear whether solving complex problems draws on different cognitive abilities or different cognitive processes than conventional intelligence test tasks.

Independently of the question whether CPS and intelligence are separate constructs, the results of study 1 revealed some practical implications for the assessment of cognitive constructs. Because intelligence and CPS were both assessed in one school lesson each, the testing times had the same duration. According to study 1, intelligence was the significantly better predictor compared to CPS and showed more substantial increments for the measures of educational success (see section 3.1). Thus, one could conclude that intelligence is a more efficient predictor for educational success than CPS. Moreover, not the administration of a complete intelligence test battery as, for example, the BIS-4 is necessary to obtain comparable results as Kretzschmar et al. (2016). Thus, when being confronted with limited testing times, intelligence should be preferred over CPS for predicting scholastic achievement.

Nevertheless, in the contexts of testing at schools, scientists sometimes struggle with acceptance difficulties when administering intelligence tests to students. Possibly, the computer-based assessment of CPS tests might be useful in such cases because high acceptance rates among students were reported for CPS tests such as MicroDYN or Genetics Lab (Greiff et al., 2012; Greiff, Wüstenberg et al., 2013; Sonnleitner et al., 2012). However, because computer-based assessments generally reveal higher acceptance rates and higher motivation levels among students (e.g.,Terzis & Economides, 2011), the computer-based administration of conventional intelligence tests might result in comparable high acceptance rates and motivation levels as the computer-based CPS assessment. Consequently, it is a task for future studies to examine whether there is a difference in students' acceptance rates and motivation levels between computer-based vs. paper-and-pencil-based CPS tests vs. intelligence tests.

In study 2, we took advantage of the often praised, but only seldom used potential of analyzing the process measures of CPS. It was revealed that the courses of strategic behaviors could be examined by latent growth curve models and students gradually increased their application of strategic behaviors if these behaviors were effective. In future studies, it would be interesting to examine whether the progressively increasing trend is generalizeable to other strategic behaviors that are effective for exploring other effect types, such as HOTAT, which is optimal for identifying interaction effects (Tschirgi, 1980). Moreover, studies concerning the learning or training of strategic behaviors might profit from the results of study 2. The increasing relative frequency course of optimal strategic behaviors in spite of increasing task difficulty gave first indications about processes that might be comparable to learning. However, as all administered MircoDYN items were structurally different, the increasing course of the optimal strategic behaviors across the task set could not be interpreted in the sense of genuine learning processes (also see section 4.3.1).

Moreover, intelligence manifested itself in facilitating the application and adaption of effective strategic behaviors. Thus, study 2 provided new insights in the mechanisms of how intelligence takes action in specific strategic actions during the exploration phase. Capturing those actions by the theory-driven definition and extraction of strategic behaviors provided information above and beyond what conventional intelligence tests could offer. Although some conventional intelligence tests could provide log-files about specific actions during the problem-solving process due to their computer-based administration, a free exploration phase in which students interact with the task is, nevertheless, unique to modern CPS tests.

Accordingly, it was assumed that the specific strategic behaviors that were applied during the unique CPS exploration phase might be the key to understand the small increment of the CPS performance measures. The results of the supplementary analysis of study 2 (see section 3.2.4) surprisingly revealed that not VOTAT but NOTAT showed unique effects on scholastic achievement in mathematics. These results seem to contradict prior research about VOTAT's importance (Bryant et al., 2015). One possible explanation – at least for the missing increment in German – might be that VOTAT's importance is highest for the science domain. As VOTAT is regarded as the key competence in scientific reasoning, this competence might be of less importance in less science-related school subjects such as verbal-domain school subjects. Although data for the domain of mathematics and language was available in the data set of Bryant et al. (2015), unfortunately, they did not analyze them. Another reason could lie in the age of the participants: whereas the students in the sample of Bryant et al.

(2015) were 11 years old when the VOTAT assessment took place, the students of this dissertation project's sample were 16 years old. As an assumption, the 11-year-old students were probably not yet taught about VOTAT, whereas the 16-year old students were presumably taught about VOTAT in science education as conducting experiments is part of the science curriculum in grades seven and eight (Ministerium für Bildung und Kultur Saarland, 2013). Thus, Bryant et al. (2015) might have assessed how young students came up with ideas on how to produce conclusive tests on their own, whereas the 16-year-old students might have referred to what they have been taught and, thus, might have merely applied their prior knowledge about variable manipulation strategies. Resorting to prior knowledge about conclusive tests might tap different cognitive processes than acquiring this knowledge on their own. This could have contributed to the different result pattern. Nevertheless, this assumption needs further clarification in future longitudinal studies with cohorts of different starting ages.

On the contrary, the surprising results about NOTAT's significance beyond intelligence seem to fit well in the line of results that indicate NOTAT's importance. Possibly, the ability of dealing with dynamic effects is of higher importance for real-life outcomes than previously assumed. Already in very early research on CPS (Dörner, 1980b; Dörner & Schaub, 1994), it was mentioned that problem solvers, who monitor autonomously developing systems, tend to be more successful in solving dynamic CPS tasks. However, regular problem solvers struggle exactly with resisting the temptation to manipulate the variables immediately and, thus, act too quickly. Comparably, more recent research also indicated the importance of NOTAT for successfully solving CPS tasks (Greiff et al., 2016) as NOTAT showed unique effects on CPS performance after controlling for VOTAT. The present results augmented the outlined current state of research with first results about NOTAT's significance for real-life outcomes. Assumingly, the small increment of CPS performance measures for educational success (compare section 3.1.3) might originate from the proficient handling of dynamic effects. However, because this is the first study to find this result and the effect size is rather small in absolute terms (1.8% of additionally explained variance for mathematics grade and only 0.4% for the mathematics competence test), it will be a task of further studies to replicate these results. Thereby, it is especially noteworthy that the latent NOTAT factor was only indicated by those four items that corresponded to the tasks with potentially dynamic effects. Thus, it would be interesting to examine whether the application of different CPS tests would lead to differential result patterns. For example, it might be plausible that CPS tests that feature dynamic effects as a more prominent characteristic (such as Dynamis2; Schoppek & Fischer, 2017) might reveal an even stronger increment of NOTAT.

Another next step in research on CPS might be to conduct studies that assess intelligence and scholastic competences via computer-based tests as well. On the one hand, this accordance in the assessment method of the predictors and the criteria of educational success would obviate possible effects of the mode of the test administration (see Sonnleitner et al., 2013). On the other hand, the exploration of parallel sequences of behaviors in CPS tests and intelligence tests would become possible. Moreover, such analyses could be complemented by time-on-task measures. For example, higher levels of intelligence might go in hand with a more efficient use of effective strategic behaviors in the sense of shorter times-on-task while being equally successful in the exploration phase.

Regarding the two key school subjects that were used as proxies for educational success, the CPS performance as well as the CPS process measures revealed a differential prediction: the incremental effects in the mathematics domain were higher than in the verbal domain. Although CPS is conceptually considered to be domain-general, study 1 and the supplementary analyses of study 2 revealed more adequate models, closer relations and higher increments in mathematics than in German. Regarding the performance measures of study 1, this pattern of result aligns well to the results of prior studies. For example, the study of Schweizer et al. (2013) reported a higher relevance of CPS for mathematics and science-related school subjects than for school subjects of the verbal domain. Moreover, in another sample, CPS predicted grades in mathematics beyond reasoning, but not grades in the native language (Kretzschmar et al., 2014). Regarding the process measures of study 2, the incremental validity of VOTAT and NOTAT in CPS contexts were examined for the first time. Also in the context of scientific reasoning, the few studies, showing VOTAT's importance for educational success, only regarded the mathematics or science domain, but not the verbal domain. Thus, no direct or meaningful comparisons about the differential relation of the process measures could be drawn.

However, when considering the demands of CPS, a strong theoretical overlap to mathematics (OECD, 2014) and the natural science subjects (Kind, 2013; Klahr & Dunbar, 1988; Wüstenberg et al., 2012) becomes obvious. For example, the mathematics problem solver has to identify and verify mathematical relations which strongly resembles the CPS knowledge acquisition phase and which is basically characterized by scientific reasoning strategies such as VOTAT or NOTAT. Furthermore, the mathematics problem solver has to interpreted, apply and evaluate the obtained results and their reasonableness which might be comparable to the CPS knowledge application phase. Thus, CPS components and solving mathematical problems seem to be theoretically corresponding and the results of this dissertation project supplied further empirical evidence for their connectedness.

Therefore, those theoretical considerations and the obtained pattern of results basically contradict the assumption of CPS being domain-general across different school subjects such as mathematics and German. The CPS measures rather appear to be content-independent if it is the task to master scientific reasoning exercises for which their importance is not questioned. Nevertheless, because scientific reasoning is not a key feature of some school subjects (as those from the verbal domain) differential relations seem to be in accordance with theory. Nevertheless, a thorough examination of the relations between the processes while solving CPS tasks, intelligence, and educational success in different domains is a promising field for further research.

To conclude, whether CPS tests really have the potential to replace conventional intelligence tests seems questionable, as intelligence revealed to be the superior predictor for educational success. Nevertheless, it is worth emphasizing that the computer-based assessment of CPS tests provides new fruitful approaches for educational researchers to investigate how the problem solving process works in detail, especially a more comprehensive examination of strategic behaviors appears to be very promising.

4.2.2 Implications for Research on Motivation in Educational Contexts

It it as major goal of educational researchers and practitioners to enhance scholastic achievement and the results of study 3 might open up new approaches for interventions specifically targeted to improves students' grades.

Because intelligence is without doubt one of the most important psychological determinants of educational success, past research repeatedly attempted to improve scholastic achievement by training cognitive abilities (for an overview see Klauer & Marx, 2010). Unfortunately, these studies revealed quite sobering results: if effects of such trainings were found at all, the very time-consuming interventions revealed only exiguous improvements (Stankov, 1986). Thus, adhering to this approach doesn't seem to be useful. As an alternative, the results of study 3 highlighted especially self-concept as a possible target of future interventions which have the goal to improve scholastic achievement. Assumingly, self-concepts are more easily influenceable compared to intelligence, as a meta-analytic review demonstrated: self-concept interventions were, for example, very effective when domain-specific self-concepts are targeted (d = 1.16; O'Mara, Marsh, Craven, & Debus, 2006). Accordingly, an attributional training which was designed to enhance academic self-concepts revealed that students attributed their success in the training to their abilities and, in turn, a lagged training effect was evident for a mathematics competence test (Dresel & Ziegler, 2006). Thus, and because self-concepts are even more important than intelligence for grades, boosting students' self-concepts could also be more efficient than cognitive trainings for improving grades. Nevertheless, the question whether targeting students' self-concepts is more effective and more efficient than training students' intelligence could only be answered in tailored intervention studies.

Concerning interest, it is especially noteworthy, that the lack of incremental effects did not imply that researchers or teachers should neglect students' interests. On the one hand, boosting students' interest might influence achievement (Hullemann & Harackiewicz, 2009) via the large amounts of common variance with self-concept. On the other hand, interests are of extraordinary importance for scholastic choices (Wigfield & Eccles, 2002). As indicated in the study by Köller, Baumert, and Schnabel (2001), academic interest became of higher importance after advanced course choices became available. The two-year longitudinal data of their study implied, for example, a substantial indirect effect of mathematics interest on subsequent mathematics achievement via its effect on course selection. The underlying mechanism might be that highly interested students tend to choose more advanced mathematics courses and, therefore, they were more likely to gain faster learning rates, which results in a deeper understanding of the mathematical topics. However, the students of this present dissertation project's sample had quite limited choices about their course selection because the students were assessed in the grade level right before making their advanced course choices. Assumingly, this might be a reason for the limited incremental validity of interest in study 3. In future studies on the relevance of academic interests for scholastic achievements, the data collection should preferably take place surrounding the advanced course choices to establish good prerequisites for the effects of interests.

Regarding the generalizability of the differential prediction pattern of cognitive and motivational variables, it seems very probable that the results are transferable to other school subjects. Although the domain-specificity is inherent in the conceptual definition of school subjects specific self-concepts and interests, domain-specific moderator effects were not reported in prior research (Wigfield & Eccles, 2002). Accordingly, prior studies numerically reported comparable differential prediction pattern of intelligence and self-concept for school subjects of the natural science domain (physics, chemistry and biology; Jansen et al., 2014) and of the verbal domain (English as second language; Zaunbauer et al., 2009). Regarding interests, a comparable numerical differential prediction pattern was obtained in four more school subjects from the verbal and the natural science-domain (German as native language, biology, chemistry, and physics; Jansen et al., 2016).

To conclude, the differential prediction pattern, indicating an extraordinary relevance of motivational variables for grades, might initiate new promising approaches for self-concept interventions that target the improvement of grades. The high probability that the differential prediction pattern is generalizable to other school subjects opens up auspicious room for versatile future intervention studies.

4.2.3 Implications for Different Achievement Indicators

In this dissertation project teacher given grades and standardized scholastic competence tests were used as indicators of educational success, because they are the two types of evidence most commonly used for decisions in educational contexts (Steinmayr et al., 2014; Willingham et al., 2002). In Germany, grades are the main criteria for evaluating educational success as they represent scholastic achievement and operate in societal and educational functions (Birkel & Tarnai, 2018; Heine, Briedis, Didi, Haase, & Trost, 2006). For example, GPA is often the single relevant achievement indicator which determines whether a student will attend academic-tracked school types or whether a student will obtain an university place. On the contrary, scholastic competence test scores are ascribed less importance in the overall institutional framework of the educational system (Füssel & Leschinsky, 2008) and are not ascribed high importance for later career or life in Germany. Thus, students might be less motivated to study extensively before they take the one-point-measurement that is mostly neither generated nor graded by the teachers. In contrast, report card grades are a conglomerate of several single grades, containing classroom examinations, oral presentations, or homework assignments and, thus, teachers might already indirectly assess effects of motivation via grades. Highly motivated students would work hardly and constantly throughout the entire school year which could result in many good single grades and, in turn, in a good report card grade.

Typically, scholastic competence tests and grades exhibit a substantial relation (e.g., $r = .40$; Marsh et al., 2005), but their correlation is far from being perfect. Thus, the question arises whether tests and grades assess the same underlying construct. Assumingly, both measures tap aspects of skill and knowledge that are generally relevant for scholastic achievement, but the achievement-relevant aspects of both measures seem to overlap only partly. With reference to standardized tests, a scholastic competence test cannot cover exactly the same material in detail that the averaged grade had assessed, even if the test is curricular valid. Moreover, the assessment format of the test could be associated with relevant or irrelevant features of achievement as teachers might stress oral presentations more than writing (Willingham et al., 2002). Nevertheless, scholastic competence tests typically show well-tested and, therefore, mostly at least sufficient psychometric properties concerning objectivity, reliability and validity (Lissmann, 2018).

With regard to grades, it is important to mention that teacher given grades are influenced by non-ability aspects, such as the teachers' subjective perceptions of the students' personality, gender, or learning behavior as well as the students' own personality and, especially, motivation (Harlen, 2005). Moreover, the psychometric properties of teacher given grades in Germany are not perfect (Birkel & Tarnai, 2018). Referring to objectivity, written examinations in mathematics as well as in German exhibit some deficiencies as different raters evaluate identical examinations with different grades (e.g., Birkel, 2003, 2005). Concerning reliability, Böhnel (1996) reported that the retest-reliability of grades was fairly high over a four year period (r = .78), but it decrease considerably with larger temporal difference between the measurement occasions. Regarding validity, the class-specific reference group effect might be the main problem of teacher given grades: students with objectively equal achievements receive different grades, depending on the general achievement level of their class (Ingenkamp, 1971). For instance, correlations between grades and standardized test scores are high within one class but considerably lower between classes (r = .70 vs. r = .30; Hoge & Coladarci, 1989; Tent, Fingerhut, & Langfeldt, 1976). In contrast, concerning another validity aspect, systematic biases in the German grading system due to irrelevant factors such as gender or socioeconomic status were found to be rather minor in elementary school (Tent et al., 1976) as well as in secondary school grades (Schrader & Helmke, 1990).

To conclude, it seems as scholastic competence tests are a purer measure of students' competences, whereas grades are more heterogeneous measures (Baumert et al., 2009; Harlen, 2005). The lower proportion of overall explained variance in grades compared to the scholastic competence tests in all analyses of the present dissertation project further evidenced this assumption. Concerning the prediction of educational success, the relevance of motivational variables was possibly underestimated in past research that did not consistently differentiate between scholastic competence tests and grades as achievement indicators. Thus, in future research, strong emphasis should be placed on distinguishing between both criteria of educational success when predicting them by cognitive and motivational variables.

4.3 Critical Acknowledgments

The following section points out possible limitations for interpreting the obtained results by critically reflecting on the instruments and the design of this dissertation project. Moreover, methodological issues of the employed analysis strategies are considered.

4.3.1 Instruments

The following section focus on the adequacy of the administered instruments for assessing intelligence, CPS, motivation, scholastic competence tests, and grades.

Assessment of Intelligence. Intelligence was assessed with a selection of 10 BIS-subtests. Because of time restrictions to one school lesson for the intelligence assessment, it was not possible to administer the complete BIS-4 test. Therefore, not all of the 12 cross classifications of the BIS-rhombus were covered. Nevertheless, the administered 10 subtests were carefully chosen. According to the results of prior studies (Brunner & Süß, 2005; Valerius & Sparfeldt, 2014), those 10 subtests were selected that revealed high loadings on their corresponding content facet. Moreover, the coverage of many content-operation-combinations was targeted. Thus, applying the criteria for a good *g* (Jensen & Weng, 1994; see section 2.1.1), it can be attested that the intelligence assessment of this dissertation project fulfilled those criteria: reliable and numerically sufficient subtests that were diverse concerning their content-operation-combinations were represented in a balanced manner within the test battery.

However, it has to be mentioned that some of the BIS cross-classifications were not covered. For example, all three BIS-cells that corresponded to the creativity operation facet were not covered. Because of the long-lasting and still ongoing debate about the definition and the reliable assessment of creativity as well as its controversially discussed relation to intelligence (e.g., Batey & Furnham, 2006), subtests of this facet were not included. Moreover, the numerical memory and the verbal speed cross-classification was not covered. Thus, only the cells that corresponded to the operation facet reasoning were fully covered as reasoning is regarded as a good marker for general intelligence (Lohmann & Lakin, 2011). However, concerning the other content and operation facets, some of the cells remained empty. Nevertheless, because the selected intelligence subtests cover a wide range of content-operation-combinations, it could be assumed that the diversity of the administered test battery was not truly endangered.

More empirical arguments for the assumption that the administered intelligence tests battery is a proper representation of *g* are relations to related constructs in expected directions and of expected magnitude. Especially reviewing the correlations with CPS performance measures is revealatory. In prior research, correlations between CPS and intelligence assessed with the short version of the BIS (Kröner et al., 2005) or the complete BIS (Kretzschmar et al., 2016; see Table 2.1) were numerically comparable to the ones obtained in this dissertation project. Moreover, they were considerably higher than the correlations between CPS and measures of figural reasoning. Thus, the maybe ”not perfect *g*” of this dissertation project was at least a ”good enough *g*” because differential effects between the broadness of the intelligence indication and its relation to CPS as well as to measures of educational success were demonstrated. In sum, it could be noted that although no complete intelligence test was assessed, the applied intelligence operationalization could be regarded as a proper representation of *g*.

Furthermore, the comprehensive intelligence assessment was an important improvement compared to prior studies in the context of CPS as well as motivation that assessed intelligence rather narrowly in the sense of only (figural) reasoning. Although intelligence test batteries that entirely consist of figural reasoning subtests, like Raven's matrices or the CFT, are without doubt good markers of intelligence (Lohmann & Lakin, 2011), these tests are not an optimal indication of a good *g*, even if they fulfill some of the criteria outlined above (Jensen & Weng, 1994, see section 2.1.1). For example, regarding the CFT, Johnson et al. (2008) reported lower correlations of the rather narrowly indicated CFT with four different (broader) intelligence test batteries and argued that it is necessary to implement some breadth in the content and format of the test battery to measure *g*. Comparably, Raven's matrices shared only approximately 50% variance with g and 10% with a fluid factor which led to the conclusion that Raven's matrices is not a pure measure of *g* (Gignac, 2015).

Empirically, it was demonstrated in study 1 (see section 3.1.4) and by Kretzschmar et al. (2016) that the incremental validity of an extracted *g*-factor was indeed affected by the composition of the test battery. Because broader measures of intelligence are assumed to explain larger amounts of variance in the criterion as rather narrow indicated measures of reasoning, other constructs such as CPS or motivation were put to a harder test compared to prior studies that used rather narrow intelligence indications. For example, in study 3 (see section 3.3), the hypothesis that motivation is a significantly stronger predictor than intelligence for grades was tested more conservatively, because of the application of broad measures of intelligence. To conclude, the comprehensive intelligence assessment corroborated the results of this dissertation project.

Assessment of CPS. For assessing CPS, MicroDYN as a reliable and frequently-used CPS assessment tool was employed (Greiff et al., 2012; Wüstenberg et al., 2012). As detailed in section 2.2.1, there is a variety of other CPS tests that could have been administered. Nevertheless, MicroDYN was chosen for several reasons. First, MicroDYN is based on minimal complex systems (Greiff & Funke, 2009). Reducing the complexity of prior CPS tests that featured up to 2000 variables (Lohhausen) to a lower boundary of complexity without being easy, made it possible to assess CPS time-economically. Thus, the administration of MicroDYN fitted well within the time constraints given by the restricted testing times in schools.

Second, MircoDYN belongs to the MCS tests. Assessing CPS by multiple complex systems comprises various advantages compared to refined classical and other modern single-item CPS tests such as Tailorshop or MultiFlux (Greiff, Fischer et al., 2015): (a) In MCS tests, there can be a variation in item difficulty.

Whereas single-item tests often contain only one difficult task, CPS tests within the MCS approach can assess CPS skills on different levels of performance by implementing multiple items with varying difficulty. (b) By assessing multiple items, adequate estimates of reliability (e.g., Cronbach's α) can be calculated and enhanced by adding more tasks to the test. Sufficient reliability is an important prerequisite for examining validity-related research questions such as the relation between CPS, intelligence, and educational success. (c) Administering multiple tasks that are independent from each other avoids the overweight of specific (and erroneous) person-task interactions. Thus, a random single misconception, which is not related to the problem solvers ability, is less likely to compound the overall CPS performance. (d) The use of multiple tasks allows the implementation of different effect types per task that, in turn, require different strategic behaviors. Therefore, the application and adaption of specific strategic behaviors (such as VOTAT or NOTAT) can be examined across the task set with changing demands.

Third, the selection of MicroDYN was oriented on recent research in the field of CPS. When establishing the test battery of the data collection, it was intended to examine the unique aspects of CPS beyond broadly operationalized intelligence (i.e. answer the research questions f study 21; see section 3.1). At that time, the latest publications on this topic investigated the incremental validity of CPS for educational success beyond measures of reasoning by using MicroDYN. To examine the unique aspects of CPS beyond broadly operationalized intelligence, the key strategy of scientific reasoning, VOTAT, implies to only vary the broadness of intelligence, but not the assessment method of CPS. Thus, to single out the effect of a broader indication of intelligence, MicroDYN as the CPS assessment method had to be held constant for drawing valid inferences.

Besides these advantages of using MicroDYN, there were also limiting aspects caused by using this CPS assessment tool. First, as MicroDYN is a representative of the minimal complex systems (Greiff & Funke, 2009), each of the administered tasks consisted of rather manageable amounts of variables and relations between the variables (Funke, 2014; Greiff et al., 2012; also see Appendix A for details about the task structures of the administered MicroDYN tasks). Thus, the question arises whether MicroDYN tasks really comply with the CPS criteria of complexity and connectedness and, in turn, whether they really measure complex problem solving. In case of MicroDYN, the tasks are designed in a way that problem solvers can completely explore the problem space (that is entirely based on linear equations) by the proper variation of the input variables (i.e. apply effective strategic behaviors such as VOTAT or NOTAT) and document the identified causal relations in a diagram in less than three minutes. Thus, once a problem solver knows how to apply the effective strategic behaviors, it is possible to explore the problem space almost routinely. This contradicts the understanding that ”there is not a single method for problem solving” (Funke, Fischer, & Holt, 2018, S. 46). In comparison to more complex CPS tests such as the Tailorshop, the contrast becomes obvious: the problem space of classical CPS tests could not be fully explored because of the higher number of variables and relations between the variables. Moreover, the problem solvers typically don't know the complete set of variables in the simulation. Accordingly, they have to rely on assumed causal relations and monitor the output variables in a more global way (Funke, 2014). Therefore, modern CPS tests like MicroDYN represent the construct of complexity and connectedness to a much smaller degree than the classical CPS scenarios.

Second, to avoid the confundation with prior knowledge, MicroDYN variables were named rather fictitiously or without deep semantic meaning. Thus, one might argue that CPS tests, which rely on minimal complex systems with abstract semantics, have forfeit their relation to everyday life. Especially in early research on CPS, intelligence tests were heavily criticized for their assumed remoteness from everyday life and intricated CPS scenarios like governing a small town appeared to be a more valid alternative to examine how people cope with the complexity of real-life. However, modern CPS tests such as MicroDYN seem to only purport closeness to everyday life. Implemented cover stories such as feeding a cat or planting pumpkins appear to be close to reality, but they only conceal that every MicroDYN task could be solved by almost routinely applying effective strategic behaviors to identify the relations between abstractly named input and output variables. Justifiably, the question arises whether modern CPS tests could still keep up the claim of being more related to real-life than traditional intelligence tests.

Summarizing the first and the second limiting aspect, it seems as CPS tests which are based on *multiple minimal* complex systems have exchanged their complexity and relatedness to everyday life for better psychometric qualities. In other words, CPS tests based on minimal complex systems ”run the danger of becoming minimal valid systems” (Dörner & Funke, 2017, S. 4). To increase the validity, it might not be useful to administer more of the same task as the MCS approach would suggest, but to include tasks with a higher variety of task requirements (Funke et al., 2018). Unfortunately, no CPS assessment tool exists yet that combines the advantages of multiple complex system CPS tests (as detailed above) with the complexity of classical CPS tests. A possible compromise might be to implement more effect types than just linear effects. A more complex problem space consisting of about five variables could easily be implemented by adding nonlinear effect types and feedback loops in – at least – some of the problems. In that way, much more observation is needed to identify the opaque relations that are unfolding over time (Dörner & Funke, 2017) which corresponds more with the genuine construct of CPS. Such 'moderate complex systems' would allow a more valid assessment of problem solving competencies within an acceptable amount of testing time (Funke et al., 2018) and with assumingly proper psychometric properties.

A third critical aspect about administering MicroDYN, specifically concerning study 1, was that using only one specific CPS test represented an item-based and not a test-based CPS assessment. Thus, one might argue that the lack of incremental validity of CPS beyond the test-based comprehensive measures of *g* (see section 3.1) merely resulted from assessing a rather ”weak CPS”. However, the item-based CPS assessment by nine MicroDYN tasks (as it was employed in this dissertation project) aligned with the psychometrically-sound and item-based CPS assessments of previous studies that revealed evidence for the incremental validity of CPS beyond narrow measures of intelligence for educational success (e.g., Greiff & Neubert, 2014: 7 tasks; Greiff, Wüstenberg et al., 2013: 7 tasks; Kretzschmar et al., 2014: 9 tasks; Wüstenberg et al., 2012: 8 tasks).

One possibility to resolve the difference between the test-based intelligence indication and the item-based CPS indication would be assessing a test-based and assumingly stronger CPS. The study of Kretzschmar et al. (2016) resolved the described difference to some extend by assessing a test-based CPS via MicroDYN and MicroFIN and revealed that even a stronger CPS could not predict educational success beyond broad measures of *g*. As a cautionary note for interpreting the results, it has to be mentioned that the intelligence assessment still comprised considerably more tests than the CPS assessment. Thus, the speculation that the pattern of results of study 1 was caused by a too weak CPS assessment cannot be fully rebutted. However, from a different point of view, it is also highly conceivable that a stronger and test-based CPS assessment might correlate even closer with intelligence and, thereby, the possibility for finding unique CPS effects beyond intelligence would be lowered even more.

Unfortunately, it remains an open question which kind of indicators of intelligence and CPS would lead to an evenness in the assessment methods of both constructs because already on the level of items, it is unclear whether a CPS item aligns to an intelligence test item. First, there is a high variability in the broadness of different intelligence items. For example, a BIS Charkow item (in which figures in a series have to be generalized and completed) is presumably broader than a X-greater item (in which numbers X greater than the prior one have to be crossed out). Second, CPS items appear to be broader than, for example, a very prototypical figural analogies item because a CPS item comprises several active interactions with the task. To conclude, further research is needed to get more insights in how a comprehensive CPS assessment would be realized that would lead to a measurement that is conceptually more comparable to the measurement of intelligence.

Shifting the focus to study 2, the assessment of nine MicroDYN tasks allowed to display the VOTAT and NOTAT courses across a sufficiently large number of tasks and especially the task type change gave the opportunity to investigate the adaption of strategic behaviors when task demands are not constant. Nevertheless, all of the nine items were administered with increasing difficulty levels and, thus, all items were structurally different (see Appendix A for details). As a consequence, final conclusions about the learning of strategic behaviors across a task set could not be drawn from study 2. Nevertheless, students increased the frequencies of their effective strategic behaviors although the tasks became more difficult. This could be interpreted as hints for processes that might be similar to learning. To validly investigate the learning of strategic behaviors, future studies are needed that compile their CPS test battery with tasks of which some subsequent tasks have the same underlying problem structure. Results of such studies could reveal valuable directions on how scientific reasoning strategies are learned and, in addition, how they might be taught by the application of CPS tests.

Assessment of Self-concept and Interest. Using the DISC (Rost et al., 2007), the measures of mathematics and German self-concept and interest were assessed by presenting the items in form of a grid. Thereby, identical item stems were administered in a very efficient way for the domains of mathematics and German. Moreover, the self-concept and interest assessment instruments revealed to be psychometrically-sound in this data set (e.g., Cronbach's $\alpha > .85$) and proofed to be valid in several prior publications (e.g., Rost & Sparfeldt, 2002; Sparfeldt et al., 2004; Sparfeldt, Schilling, Rost, & Thiel, 2006).

However, as a cautionary note, it has to be mentioned that there was a difference between the number of administered self-concept items and the number of administered interest items in the questionnaires. Whereas self-concept was assessed with eight items, interest was assessed with only four items. Thus, one could argue that there might have been an unevenness in the indication of self-concept and interest which has caused the lack of incremental effects of interest when controlling for the shared variance with self-concept as reported in study 3 (see section 3.3). Speculatively, a more equivalent indication of self-concept and interest in the sense of either less self-concept items or more interest items possibly would have led to a stronger impact of interest. However, academic self-concept as well as interest are highly homogeneous constructs. Therefore, a reduction of the number of self-concept items should not lead to a narrower self-concept factor that is less predictive for the criteria. In turn, an increase of the number of interest items should not lead to a broader interest factor that is more predictive for the criteria. Thus, it appears to be unlikely that a different indication would have caused a different pattern of results. Additional analyses that reduced the number of the self-concept items (not reported in detail) tested this conjecture and revealed virtually no changes in the result pattern of study 3 (Median $|\Delta\beta|$ = .02). Thus, those additional analyses rather substantiated the results of study 3, by revealing that the result pattern was not caused by a potential unevenness in the factor indications and that an allegedly weaker indicated self-concept factor still has the power to predict educational success beyond interest and comprehensive measures of intelligence.

Assessment of Scholastic Competence Tests. The basis of the administered scholastic competence tests were the mathematics competence test and the reading comprehension test of the KESS 10/11 study (Vieluf et al., 2011). As mentioned in section 3.1.2, the competence tests relied on selected KESS 10/11 items. Items in mathematics were carefully selected so that the key mathematical concepts were still adequately represented. The reading comprehension competence test was composed by three selected tests that fulfilled the criteria of being of comparable length and covering different domains (narrative text, expository text, newspaper article).

Regarding reliability, values of Conbach's α for the administered mathematics competence test and the reading comprehension test were comparable to the values reported in the KESS 10/11 study. Concerning validity, the rationale behind assembling the original KESS items and texts is well documented in the KESS 10/11 publication (Vieluf et al., 2011). For instance, the mathematics test focused on critical mathematical concepts like algebra, analytic geometry, trigonometry, linear and quadratic relations, and stochastic as they are documented in the reference framework for mathematics of the federal state institute for teacher education and school development in Hamburg. The curricular validity of the tasks was assured by an expert group of the federal state institute for teacher education and school development in Hamburg. The same applies to the reading comprehension test.

Moreover, the KESS 10/11 publications reported several relations between the performance in the scholastic competence tests and other variables that indicated convergent and divergent validity. For example, the performance correlated with school type (students from academic-tracked schools scored higher than students from non-academic-tracked schools), with language spoken at home (German speaking students scored higher than non-German speaking students), with educational attainment of the parents (students of parents with higher educational levels scored higher than students of parents with lower educational levels), and with social class (students of higher social classes scored higher than students of lower social classes). Differential correlations were obtained for gender: boys scored higher than girls in the mathematics competence test, but girls scored higher than boys in the reading comprehension test. These relations were in expected directions and, thus, gave hints for the tests' validity. The 30 selected mathematics items and the three selected reading comprehension texts used in this dissertation project revealed similar relations to gender and language spoken at home (the other variables mentioned above were not assessed). Therefore, these comparable relations support the validity of the mathematics and the reading comprehension competence test in the sample of this dissertation project.

As an aspect that has to be critically referred to, it has to be mentioned that the scholastic competence test in the German subsample specifically focused on reading comprehension and not on the school subject German in general. In contrast, the items to assess German self-concept and interest focused on the school subject German as a whole. Thus, the items to assess motivation in German and the reading comprehension test did not align perfectly to each other. Although reading comprehension is a central aspect of the school subject German (Cain & Oakhill, 2006), the administered reading comprehension test did not tap the central aspects of the school subject German in 10th and 11th grade as a whole. Nevertheless, at least the three administered reading comprehension texts featured a variety of genres such as narrative, expository, and newspaper article as well as a variety of comprehension levels such as text-based questions, local inferences, and global inferences. Thus, the reading comprehension test covered at least partially some of the skills that are part of the curriculum. Nevertheless, it has to be clearly stated that the administered reading comprehension test was not curricular valid for students from upper secondary education grade levels. Thus, the effects of the non-perfect correspondence between the motivational measures and the achievement measures prevented the meaningful analyses of the German subsample. Possibly, the unacceptable fit statistics of the German subsample as described in the supplementary analyses of study 3 (see section 3.3.4) also resulted from this misalignment.

Assessment of Grades. Inquiring school grades of students is a viable way to assess educational success. They represent scholastic achievement and exhibit a profound predictive validity over longer time periods (Birkel & Tarnai, 2018). Especially, self-reported grades are widely-used because of their easy accessibility. Nevertheless, relying on self-reported school grades could also be criticized for some reasons. For example, in meta-analytic results, inaccuracies and biases in self-reported GPA were evidenced (Kuncel, Credé, & Thomas, 2005). However, in contrast to Kuncel et al. (2005), the half term report card grades in two particular school subjects and not final high school GPA or college GPA were assessed. Therefore, these findings are not completely applicable to the present data.

Moreover, meta-analyses have the potential limitation to reflect certain populations to a lesser degree than other populations. Because most of the studies that were included in the meta-analysis of Kuncel et al. (2005) stemmed from Anglo-

American countries, including high school students and college students, they might not be generalizeable to German high school students. Further evidence for this assumption can be gathered by reviewing two studies that investigated the accuracy between self-reported and actual report card grades in particular school subjects for German high school students. For example, Sparfeldt, Buch, Rost, and Lehmann (2008) reported a correlation between actual and self-reported grades in mathematics of r = .94 and that 91% of the tenth-grade-students indicated their grade correctly (difference between self-reported and actual grades: d = 0.07; 1% under-estimators, 8% over-estimators). Comparably, Dickhäuser and Plenter (2005) found a correlation of r = .88 between actual and self-reported grades in mathematics for eighth-grade-students. Thus, both studies concluded that self-reported grades seem to be very accurate indicators of actual grades. Moreover, at least in mathematics, the level of accuracy in reported grades was not influenced by the students' competence levels (Dickhäuser & Plenter, 2005). Sparfeldt et al. (2008) evidenced a very small effect in the sense of a slightly higher over-estimation of lower-achieving students (η^2 = .005).

In sum, it seems that self-reported grades are very appropriate and efficient measures of actual grades – at least in research with German high school students of the age of the present sample. Thus, effects that impair the interpretation of the overall findings of this dissertation project due to inaccuracies or biases in self-reported grades seem to be unlikely.

4.3.2 Design of the Dissertation Project

In the following section, limiting aspects due to the design of this dissertation project are considered. Specifically, the assessment of cross-sectional data, the sample size, splitting up the sample in two halves, and the generalizability of the results to other populations are considered in more detail.

Cross-sectional Data. As a potential limiting factor, it has to be mentioned that the data of this dissertation project was cross-sectional. Thus, the data provided information about the variables' interplay at a certain point of time and about the amounts of unique variances, but the data could not be interpreted as evidence for causal ordering.

Especially with regard to study 3 (see section 3.3), it has to mentioned that self-concepts, interests, and scholastic achievements are mutually dependent (e.g., Marsh et al., 2005). For example, referring to self-concept, prior research contrasted the self-enhancement model, in which higher self-concepts result in higher achievement, with the skill-development model, in which higher achievements result in higher self-concepts. It was revealed that the effects from achievement on self-concept appeared to be larger than those from self-concept on achievement, favoring the skill-development model (Marsh & O'Mara, 2009). Moreover, there are also strong hints for a reciprocal dependency of self-concept and interest, which evidenced that students come to value the school subjects they think they are good at across the years and, in addition, that students think they are good at in the school subjects they value (Marsh et al., 2005).

In conclusion, gathering longitudinal data about the relations of cognitive, strategic-behavioral, and motivational variables with educational success would have been desirable to provide insights in the causal ordering of the variables. However, because of the cross-sectional design, this dissertation project focused on the statistical prediction of educational success in the sense of explained variances.

Sample. In this dissertation project, data of *N* = 496 German (federal state: Saarland) high school students that stemmed from two academic-tracked school types (Gymnasium and Gemeinschaftsschule) were assessed. As potentially limiting aspects, the sample size and the possible effects of only having assessed students from academic tracks need to be discussed in more detail.

Concerning the sample size, no prior planning of sample sizes with conventional software packages (e.g., G*Power) was possible because of the hierarchical data structure (students in classes) and the nature of some research questions (testing path coefficients against each other instead of against zero; Lee, Cai, & MacCallum, 2012). Therefore, we relied on the simulation study of Wolf, Harrington, Clark, and Miller (2013) which could be used as a rough proxy for the planning of sample sizes for computing structural equation models. According to their study, the recommended sample size for models with three latent factors (with 3 or 4 indicators each), factor loading of $\lambda \leq .65$, and a power of $\beta \leq .80$ was $N \approx 220$ participants. In this dissertation project, the analyzed statistical models comprised latent factors that were mostly indicated by more than four indicators. Thus, the sample size of $N \approx 220$ could be regarded as an upper boundary because minimum sample size requirements decrease as the number of indicators increases (Wolf et al., 2013). On the basis of these considerations, it was aimed for a minimum sample size of $N \approx 200$ - 250 students when planning the data collection. This was accomplished. Analyses with the smallest sample size were conducted in study 1 and 3 (mathematics subsample, see sections 3.1.2 and 3.3.2). For these analyses, the sample comprised *N* = 245 students each which is still in the upper area of the intended range. Thus, it appears to be unlikely that the conducted analyses suffered from insufficient statistical power.

With regard to the restrictions due to only having assessed academic-tracked students, it has to be mentioned that some of the assessed variables might have reduced variances, whereas others don't: The variances of intelligence, CPS, and the scholastic competence tests might be reduced in the present academic-tracked sample compared to a more representative sample, whereas the variances of self-concept, interest, and grades are typically not reduced in academic-tracked samples (for self-concept see Rost et al., 2007).

Typically, more than 50% of a cohort attend academic-tracked school types in Germany. In Saarland, the federal state where the present data stemmed from, this percentage is even higher. In total, around 60% of the students in Saarland finish the academic-tracked schools by obtaining the general matriculation standard or the advanced technical college certificate (Malecki, 2016). Thus, although this sample only comprised academic-tracked schools, it could be assumed that the sample exhibits a certain degree of heterogeneity. Inspecting the high correlations between intelligence and CPS in study 1 ($r_{\text{mathematics/German subsample}}$ = .69/.76) as well as between intelligence and the mathematics competence tests (r = .85), their high magnitude generally speaks against severe variance reductions. Moreover, in study 3, the difference between the intelligence and motivation path coefficients revealed to be considerably high. Assumingly, even in an unselected sample (which would go hand in hand with an increase in the paths related to intelligence and competence tests), it would be very conceivable that the substantial numerical differences remain statistically significant.

In summary, the sample of this dissertation project had a sufficient sample size to properly answer the research questions. Nevertheless, potential variance reductions due to the academic-tracked sample might be present. However, as the conducted studies did not aim to establish norms or benchmarks, the sample under investigation seems fairly adequate to answer the research questions.

Sample Split up in Two Halves. When assessing the data of this dissertation project, the testing time was limited to three consecutive school lessons due to the practical restrictions of testing in schools. Because the assessment of intelligence and CPS each took a whole lesson, there was only one lesson left for administering the scholastic competence tests. As a consequence, the sample was split up randomly into two halves so that each half worked on either the mathematics competence test or on the reading comprehension test. Thus, a sufficient number of scholastic competence test items to establish good prerequisites for a reliable measurement could be administered for both school subjects within one school lesson. Unfortunately, the resulting two subsamples made it impossible to examine the relations between the mathematics competence test and the reading comprehension test[4]. Indubitably, it would have been desirable to have obtained data of both scholastic competence tests from each student. However, a total

[4] The mathematics and German grade were assessed for the full sample; their manifest correlation revealed to be r = .33; $p < .05$.

assessment time of four lessons would have been necessary for this intention. Unfortunately, extending the total testing time to four lessons would have led to further issues: Assessing students competencies and attitudes in four consecutive lessons might result in severe strains that would have negatively affected the quality of the assessed data. Otherwise, spreading the four lessons across two different assessment dates might have gone along with the resistance of teachers and principals and with missing values because of ill or absent students to one of the two dates. Therefore, randomly splitting up the sample in two subsample appeared to be the most efficient and feasible compromise to collect the broad data set.

Generalizability to Other Populations. The results of this dissertation project revealed insights in how educational success is predicted in a sample of German high school students. Thus, the question arises whether the findings could be generalized to other populations. Answering this question, differential conclusions had to be drawn for the studies 1 and 2, including the measures of CPS, and the study 3, including measures of motivation.

With reference to the studies 1 and 2 (see section 3.1 and 3.2), it seems conceivable that the results are generalizable to other German students. Although, as mentioned in this section, the sample under investigation might have reduced variances in the variables intelligence, CPS, and scholastic competence tests, such possible variance reductions should have rather small effects on the overall result pattern. However, possible effects could have been lowered correlations between intelligence and the performance as well as process measures of CPS. Specifically in study 1, the examination of a more representative sample, consisting of students from all school types, might have resulted in even higher correlations between intelligence and the CPS performance scores and, in turn, in an even lower probability for CPS to incrementally explain variances in the scholastic competence tests or grades. Thus, it is not plausible to expect higher CPS increments in more representative samples. Addressing study 2, also higher correlations between intelligence and the CPS process measures might have resulted from examining more heterogeneous samples. This would further foster the result that intelligence manifests itself in the application of effective strategic behaviors.

With reference to students from different countries, it appears plausible to not expect substantially different pattern of results for studies 1 and 2. As a first cross-country study about MicroDYN has shown, MicroDYN items exhibit measurement invariance across students from Germany and Hungary (Wüstenberg, Greiff, Molnár, & Funke, 2014). This result gave first indications that merely the skill level might vary between the countries, but not the way MicroDYN is measuring complex problem solving abilities.

Concerning the measures of motivation, it appears to be highly probable that the differential prediction pattern of study 3 (intelligence is the better predictor for achievement tests, whereas self-concept is the better predictor for grades, see section 3.3) might be generalizable to other German students. Across all federal states of Germany, the overall institutional framework of the educational system is quite similar with regard to the high importance of grades and the rather low importance of scholastic competence tests (Füssel & Leschinsky, 2008). However, when regarding students from other nations, the differential prediction pattern might emerge less clearly because motivation might play a more important role for predicting scholastic competence tests. In Germany, grades often represent the single relevant achievement indicator, whereas scholastic competence tests are not ascribed high importance (Birkel & Tarnai, 2018). In contrast, in other nations, scholastic competence tests are of much higher significance and it is likely that students purposefully prepare themselves more before taking these kinds of tests (e.g., Hansen, 2004; Powers & Rock, 1999). Therefore, motivational variables and, especially self-concept, might reveal coefficients that are numerically closer to those of intelligence. In this line of argumentation, it seems conceivable that the differential prediction pattern might emerge less clearly for scholastic competence tests and that the unique aspects of self-concept might reveal a higher increment on scholastic competence tests beyond intelligence. However, although several prior studies revealed solid evidence for the differential prediction pattern, the majority of these samples only comprised German students and not international students (Helmke, 1992; Jansen et al., 2016, 2014; Marsh et al., 2005; Steinmayr & Meißner, 2013). Thus, it is still a task for future studies to examine the differential prediction pattern in an international sample. At least, a numerically higher relevance of self-concept for grades than for tests is internationally underpinned by the meta-analytic findings of Möller et al. (2009).

Regarding students of different ages, the generalizabilty of the differential prediction pattern of study 3 might not be given for younger students. Typically, the relations between intelligence and grades seem to decrease with age (Jensen, 2012), whereas motivation shows an increasing relation to grades because self-concepts and interest become more differentiated and stable with age (Skaalvik & Valås, 1999). Therefore, the impact of motivation on achievement-related behavior is possibly not yet fully developed in younger students. Accordingly, empirical evidence for the comparable importance of cognitive and motivational variables in elementary school was revealed for the school subjects mathematics and German (Schneider, Lotz, & Sparfeldt, 2018; Weber, Lu, Shi, & Spinath, 2013).

Discussing the generalizability of the present sample more generally, the participation in the data collection partly depended on the decision of the students' schools principal and teachers as well as their parents. Therefore, the selection of the students was not totally at random. Nevertheless, the high participation rate within the sample made a substantial bias due to systematic self-selection unlikely.

4.3.3 Methodological Issues

In this dissertation project, a variety of advanced and sophisticated methods that were specifically tailored to the nature of the research questions were applied. The following section provides a critical reflection on the employed analysis strategies by considering the key aspects of comparing standardized paths, examining unique aspects, and establishing latent growth curve models. Furthermore, the trustworthiness of model fit indices are discussed.

Comparisons of Standardized Paths. When aiming to statistically compare standardized paths, it is a general obstacle that common statistical software packages have no built-in function for comparing standardized path coefficients. Instead, it is only possible to statistically compare non-standardized paths by constraining them to be equal and subsequently performing likelihood ratio tests. However, those non-standardized paths typically don't have the same metric and, therefore, they cannot be compared meaningfully. Thus, advanced statistical methods have to be employed to resolve the metric problem. Accordingly, studies 1 and 3 employed two slightly different approaches for comparing standardized path coefficients that were specifically matched to the particular research questions and the established statistical models.

In study 1, we compared those two paths that connected two independent latent variables with one depended variable. For instance, the path from intelligence on grades and the path from CPS on grades would be such paths. If the latent factors of intelligence and CPS were assigned the same factor variance, the metric of their paths on grades would be identical. Thus, it was adequate to constrain their non-standardized paths (after they have been assigned the same metric) to be equal to produce conclusive tests of their β-weights. Differences between paths were subsequently tested by χ^2-difference tests (see Appendix B for details).

In study 3, the procedure of study 1 could not be applied. Not only paths from two independent latent variables on one dependent variable were compared (e.g., comparing the path from intelligence on grades with the path of self-concept on grades), but also paths from one independent variable on two dependent variables of which one variable is manifest. For example, the path from intelligence on grades and the path from intelligence on the scholastic competence test would be such paths. As the grades were specified as a manifest variable, no latent factor variance could be altered. Consequently, the path from intelligence on grades and the path from intelligence on the scholastic competence test cannot be assigned the same metric by using the convenient method of study 1. For comparing the two paths from self-concept on both dependent variables or from interest on both dependent variables, the same problem applies. Thus, study 3 used a more advanced alternative method and relied on reparamerization (Kwan & Chan, 2011). This method transforms the original model into a covariance-equivalent transformed model. In this transformed model, the non-standardized path coefficients are equal to the standardized path coefficients of the original model. Thus, it was possible to conduct meaningful and interpretable Wald tests of path differences by common statistics software such as M*plus* (see Appendix B for details).

To conclude, the methodological procedures of studies 1 and 3 made it possible to compare standardized path coefficients of different variables that were incommensurable before because of their different metrics. By the application of these state-of-the-art methods, meaningful and reliable answers to the path-comparison-related research questions of studies 1 and 3 were given.

Examination of Unique Variances. As a main research aim of this dissertation project, unique variance proportions of intelligence, motivation, and CPS performance as well as process measures for predicting educational success were examined. Thereby, residual models and the Cholesky factoring approach (de Jong, 1999; Loehlin, 1996) were utilized. Basically, both approaches resemble hierarchical regression analyses, but they are more efficient because the examination of increments could be conducted within only one model for each increment. Moreover, the latent analysis of unique effects could be conducted without altering the model fit or affecting the measurement part of the model. As another virtue, squared path coefficients could be straightforwardly interpreted in the sense of explained variances.

Contrasting residual models and the Cholesky factoring approach, the number of predictors has to be considered. When having only two predictors in the latent model, residual models are more parsimonious because one phantom variable less than in a Cholesky factoring model has to be specified. In contrast, when having three or more predictors in the model, the Cholesky factoring approach has to be preferred because the assignment of priorities to the predictors makes it more lucid and transparent which phantom variable represents which shared and which unique variance proportions. Accordingly, in the analyses of study 1 and the supplementary analyses of study 2, the residual models were used because they featured only two-predictor-models. In the analyses of study 3, which featured three-predictor models, the Cholesky factoring approach was applied.

Latent Growth Curve Modeling. In study 2, the course of the strategic behaviors VOTAT and NOTAT across a CPS task set with changing demands was analyzed by means of LGCM (see section 3.2). LGCMs are well-suited to examine the change of a variable over time in a structural equation framework. Thereby, they consist of an intercept factor that represents the initial or the reference level of the growth curve and a slope factor that represents the growth trajectory of the curve (Hancock et al., 2013). Typically, such models estimate smoothed trajectories and tend to fail when sharp changes are present in the data. However, the assessment of MicroDYN featured exactly such sharp changes. More specifically, the task type change from non-dynamic to dynamic tasks and the additional instruction after task five could be regarded as an event that caused a discontinuity in the growth curve and that affected the slope as well as the intercept factor. Thus, the analyses in study 2 were based on a discontinuous or piece-wise LGCM design (Diallo & Morin, 2015). Such designs are characterized by modeling more than just one slope or intercept factor to adequately represent the growth trajectory before and after an event that caused the discontinuity. Moreover, such designs allow the intercept and slope factors to differentially correlate with other variables. Thus, the modeling of a discontinuous LGCM with two intercept and two slope factors, which represented the levels and the growth trajectories before and after the task type change in the CPS task set, seemed especially well-suited to analyze the courses of VOTAT and NOTAT and to examine their differential relations to intelligence.

Moreover, in study 2, it was explicitly tested whether the hypothesized discontinuous (vs. continuous) growth curve and the assumed linear (vs. quadratic and vs. no-change) slope represented the data structure adequately. Thus, methodologically very good prerequisites for an appropriate representation of the course of the strategic behaviors and their relations to intelligence were established.

As a cautionary note, it has to be mentioned that VOTAT as well as NOTAT were measured only with one indicator per CPS task. Thus, no second-order LGCMs that adjust for measurement errors could be specified. As a consequence, insufficient reliability might have negatively affected the estimation of the courses. However, the good and acceptable fit indices for both models indicated that the data was properly represented by the discontinuous LGCMs. Further evidence was provided by the manifest inter-task-correlations between two adjacent tasks. Regarding VOTAT, the median was *Mdn(r)* = .73 across the first five tasks and *Mdn(r)* = .86 across the last four tasks. Regarding NOTAT, the median was *Mdn(r)* = .42 across the first five tasks and *Mdn(r)* = .58 across the last four tasks. These values indicated that students did not just randomly showed a specific strategic behavior. In sum, these results were reasonable markers for the reliability and the validity of the indicators and, consequently, the results of study 2.

Fit Indices. The results provided in the studies 1 – 3 were all based on at least acceptable fitting models. However, some of the models reported in the supplementary analyses fitted mediocrely or worse (Little, 2013). Results of these models were not reported in detail, because they are hardly interpretable as the estimates might be biased (Schermelleh-Engel et al., 2003; West, Taylor, & Wu, 2012).

The reasonableness of relying on the cut-off values as specified by, for example, Hu and Bentler (1999) is not undisputed and recently questioned again by Greiff and Heene (2017). Because goodness of fit indicators depend on a number of factors that are unrelated to actual model fit, no 'golden rules' but only 'rules of thumbs' about their applicability exist. Thus, the main risk of adherently relying on strict cut-off values is that correctly specified models could be wrongly rejected and, the other way round, that misspecified models could be wrongly accepted. In the light of this debate, the question arises how trustworthy the goodness of fit indices really are. Thus, Greiff and Heene (2017) recommended to rely on inspecting local misspecifications. For inspecting local misspecifications, Schermelleh-Engel et al. (2003) recommend to consider the largest modification indices (MIs). Nevertheless, model modifications based on MIs should be defensible from a theoretical point of view because MIs are highly susceptible to capitalization on chance which can cause a lack of model validity (MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992).

Accordingly, the MIs of the unacceptable fitting models of the supplementary analyses of the studies 2 and 3 were inspected. For example, the largest MIs of the residual NOTAT-model for the German subsample of study 2 concerned correlations among the first-order intelligence factors. Allowing these factors to correlate doesn't make sense because the shared variance among the first-order factors was captured in the second-order *g*-factor. In the supplementary models of the German subsample of study 3, the by far largest modification indices pertained residual correlations among some of the self-concept items and among some of the interest items. Unfortunately, the highest MIs of the three models did not correspond across the models. Therefore, such model modifications could not be fostered by theoretical considerations and were not implemented.

Summarily, there are no unambiguous goodness of fit guidelines how to handle the non-acceptable fitting models of the supplementary analyses. Moreover, the recommended MI inspection revealed inconclusive results. As a consequence, the results of the unacceptable fitting models were not reported in detail because falsely accepting erroneous models is more serious than the failure to reject correct models (Greiff & Heene, 2017).

4.4 Final Conclusion

In this dissertation project, it was aimed to examine the incremental validity of selected cognitive, strategic-behavioral, and motivational variables beyond broadly operationalized intelligence on educational success. To pursue this aim, three separate empirical studies were conducted that revealed new insights in the interplay of intelligence and the examined variables when predicting scholastic achievement by scholastic competence tests and grades in mathematics and German.

In Study 1, the incorporated comprehensive intelligence assessment revealed a large overlap between CPS and *g*. Moreover, intelligence proofed to be a stronger predictor than CPS for educational success. In addition, whereas CPS only provided a substantial increment above and beyond broadly assessed intelligence on the mathematics competence tests, intelligence revealed substantial increment on all considered indicators of scholastic achievement. Thus, intelligence prevailed as the superior predictor for educational success compared to CPS. In light of these findings, it appears questionable whether CPS assessments fulfill the probably exaggerated expectations to complement or even replace conventional intelligence tests in educational settings.

The actual potential of CPS tests might rather lie in their interactive exploration phase and its computer-based assessment that provides detailed information about the problem solving process. Study 2 made use of this repeatedly praised potential by the theory-driven definition of the optimal strategic behaviors VOTAT and NOTAT and their subsequent extraction from the corresponding log-files. Modeling the course of the strategic behaviors across a task set with changing demands revealed that students used those strategic behaviors more frequently which were effective and that the students were able to flexibly adapt their strategic behaviors to occurring task type changes. Additionally, more intelligent students applied the effective strategic behaviors with higher frequencies and partly adapted them with a steeper gradient. Thereby, study 2 showed that intelligence facilitated the problem solving process. Additional analyses focused on the incremental effects of the strategic behaviors for predicting educational success and revealed rather sobering results: whereas VOTAT showed no substantial unique effects, NOTAT exhibited only very small unique effects beyond intelligence on scholastic achievement in mathematics. Thus, the potential of the CPS process measures rather seem to lie in the possibility to investigate the problem solving processes in more detail than in predicting educational success beyond intelligence.

Shifting the focus towards the interplay between intelligence and motivational variables, study 3 clarified that intelligence prevails as the more important predictor for scholastic competence test, but that subject-specific academic self-concept and interest revealed to be of higher relevance than intelligence when predicting grades. However, when the shared variance between both motivational variables was considered, only the differential prediction pattern of self-concept remained substantial. This indicated that the formerly substantial prediction pattern of interest on grades was basically caused by the large overlap between self-concept and interest. Nevertheless, as self-concept showed large unique effects on grades beyond broad measures of intelligence, the attention of researchers and practitioners, who aim to improve scholastic achievement, might be shifted towards new approaches that specifically focus on improving students self-concepts in intervention studies, aiming to increase students grades.

In sum, the results of this dissertation project emphasized intelligence as one of the most important psychological predictors of educational success. Other cognitive or strategic-behavioral variables such as performance and process measures of CPS contributed rather negligibly to the prediction of educational success if intelligence is indicated by broad and comprehensive measures. In contrast, motivational variables and intelligence have far less overlap, which ensured better prerequisites for a substantial prediction pattern. Accordingly, it was revealed that self-concept was the strongest predictor for grades among intelligence and interest. Thereby, the extraordinary importance of self-concept as well as the importance of distinguishing between different indicators of scholastic achievement when predicting educational success was emphasized.

References

Adamson, S. L., Banks, D., Burtch, M., Cox, F., Judson, E., Turley, J. B., … Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching*, *40*, 939-957. doi: 10.1002/tea.10117

Batey, M., & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, Social, and General Psychology Monographs*, *132*, 355-429. doi: 10.3200/mono.132.4.355-430

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, *4*, 165-176. doi: 10.1016/j.edurev.2009.04.002

Benedek, M., Jauk, E., Sommer, M., Arendasy, M., & Neubauer, A. C. (2014). Intelligence, creativity, and cognitive control: The common and differential involvements of executive functions in intelligence and creativity. *Intelligence*, *46*, 73-83. doi: 10.1016/j.intell.2014.05.007

Böhnel, E. (1996). Die Frage der Prognostizierbarkeit von Schulerfolg in der Sekundarstufe I aufgrund der Benotung in der Primarstufe [Prognosis of school achievement by the results in primary school]. *Unterrichtswissenschaft*, *24*, 343-360.

Birkel, P. (2003). Aufsatzbeurteilung – ein altes Problem neu untersucht [Assessment of essays - new examination of an old problem]. *Didaktik Deutsch*, *9*, 46-53.

Birkel, P. (2005). Beurteilungsübereinstimmung bei Mathematikarbeiten? [Evaluation agreement in mathematics examinations]. *Journal für Mathematik-Didaktik*, *26*, 28-47. doi: 10.1007/bf03339005

Birkel, P., & Tarnai, C. (2018). Zensuren und verbale Schulleistungsbeurteilung [Grades and verbal evaluations of scholastic achievement]. In D. H. Rost, J. R. Sparfeldt, & S. R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie* (5th ed., p. 904-917). Weinheim, Germany: Beltz.

Bitner, B. L. (1991). Formal operational reasoning modes: Predictors of critical thinking abilities and grades assigned by teachers in science and mathematics for students in grades nine through twelve. *Journal of Research in Science Teaching*, *28*, 265-274.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* New York, NY: Wiley. doi: 10.1002/0471746096

Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist*, *34*, 139-153. doi: 10.1207/s15326985ep3403_1

Brehmer, B. (1987). Development of mental models for decision in technological systems. In J. Rasmussen, K. Duncan, & J. Leplat (Eds.), *New technology and human error* (p. 111-120). Chichester, UK: Wiley.

Brühwiler, C., & Helmke, A. (2018). Determinanten der Schulleistung [Determinants of scholastic achievement]. In D. H. Rost, J. R. Sparfeldt, & S. R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie* (5th ed., p. 78-91). Weinheim, Germany: Beltz.

Brody, N. (2000). Intelligence. In A. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. 4, p. 318-324). Washington, DC: American Psychological Association.

Brunner, M. (2008). No *g* in education? *Learning and Individual Differences*, *18*, 152-165. doi: 10.1016/j.lindif.2007.08.005

Brunner, M., & Süß, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Education and Psychological Measurement*, *65*, 227-240. doi: 10.1177/0013164404268669

Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2015). The importance of being able to deal with variables in learning science. *International Journal of Science and Mathematics Education*, *13*, 145-163. doi: 10.1007/s10763-013-9469-x

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*, 683-696. doi: 10.1348/000709905x67610

Calvin, C. M., Fernandes, C., Smith, P., Visscher, P. M., & Deary, I. J. (2010). Sex, intelligence, and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in england. *Intelligence*, *38*, 424-432. doi: 10.1016/j.intell.2010.04.005

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (p. 5-21). Amsterdam, Netherlands: Pergamon.

Chamorro-Premuzic, T., & Furnham, A. (2008). Personality, intelligence, and approaches to learning as predictors of academic performance. *Personality* *and Individual Differences*, *44*, 1596-1603. doi: 10.1016/j.paid.2008.01.003

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098-1120. doi: 10.1111/1467-8624.00081

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Psychology Press. doi: 10.4324/9780203771587

Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ: A latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence*, *39*, 323-334. doi: 10.1016/j.intell.2011.06.004

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13-21. doi: 10.1016/j.intell.2006.02.001

de Jong, P. E. (1999). Hierarchical regression analysis in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 198-211. doi: 10.1080/10705519909540128

Deák, G. O. (2003). The development of cognitive flexibility and language abilities. *Advances in Child Development and Behavior*, *31*, 271-327. doi: 10.1016/s0065-2407(03)31007-9

Diallo, T. M. O., & Morin, A. J. S. (2015). Power of latent growth curve models to detect piecewise linear trajectories. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 449-460. doi: 10.1080/10705511.2014.935678

Dickhäuser, O., & Plenter, I. (2005). Zur Akkuratheit selbstberichteter Noten [On the accuracy of self-reported school marks]. *Zeitschrift für Pädagogische Psychologie*, *19*, 219-224.

Dresel, M., & Ziegler, A. (2006). Langfristige Förderung von Fähigkeitsselbstkonzept und impliziter Fähigkeitstheorie durch computerbasiertes attributionales Feedback [Long-term enhancement of academic self-concept and implicit ability theory through computer-based attributional feedback]. *Zeitschrift für Pädagogische Psychologie*, *20*, 49-63. doi: 10.1024/1010-0652.20.12.49

Dörner, D. (1979). Programm TAILORSHOP in der Version für TI-59 mit Drucker PC-100. Modifizierte und kommentierte Fassung von Norbert Streitz. [Computer software manual].

Dörner, D. (1980a). Heuristics and cognition in complex systems. In R. Groner, M. Groner, & W. F. Bischof (Eds.), *Methods of heuristics* (p. 98-108). Hillsdale, NJ: Lawrence Erlbaum.

Dörner, D. (1980b). On the difficulties people have in dealing with complexity. *Simulation & Gaming*, *11*, 87-106. doi: 10.1177/104687818001100108

Dörner, D. (1986). Diagnostik der operativen Intelligenz [Assessment of operative intelligence]. *Diagnostica*, *32*, 290-308.

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology*, *8:1153.* doi: 10.3389/fpsyg.2017.01153

Dörner, D., Kreuzig, H. W., Reither, F., & Strohschneider, S. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität [Lohhausen. On dealing with uncertainty and complexity].* Bern, Switzerland: Huber.

Dörner, D., & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. *Applied Psychology*, *43*, 433-453. doi: 10.1111/j.1464-0597.1994.tb00839.x

Dörner, D., Stäudel, T., & Strohschneider, S. (1986). *Moro: Programmdokumentation [Moro: Program documentation] (Memorandum No. 23).* Bamberg, Germany: University of Bamberg, LS Psychologie II.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczale, C. M., Meece, J. L., & et al. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (p. 75-146). San Francisco, CA: Freeman.

Fischer, A., Greiff, S., & Funke, J. (2017). The history of complex problem solving. In B. Csapó & J. Funke (Eds.), *The nature of problem solving. Using research to inspire 21st century learning* (p. 107-121). Paris, France: OECD Publishing. doi: 10.1787/9789264273955-en

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* (p. 3-25). Hillsdale, NJ: Erlbaum.

Freudenthaler, H. H., Spinath, B., & Neubauer, A. C. (2008). Predicting school achievement in boys and girls. *European Journal of Personality*, *22*, 231-245. doi: 10.1002/per.678

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, *15*, 373-378. doi: 10.1111/j.0956-7976.2004.00687.x

Füssel, H.-P., & Leschinsky, A. (2008). Der institutionelle Rahmen des Bildungswesens [The institutional framework of the education system]. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer, & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick* (p. 131-203). Reinbeck, Germany, Rowohlt.

Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? [Some remarks on the problems of problem solving research or: Does test intelligence predict control performance?]. *Diagnostica*, *29*, 283-302.

Funke, J. (1984). Diagnose der westdeutschen Problemlöseforschung in Form eigener Thesen [Assessment of West German problem solving research]. *Sprache & Kognition*, *3*, 113-129.

Funke, J. (2001). Dynamic systems as tools for analyzing human judgement.

Thinking and Reasoning, *7*, 69-89.

Funke, J. (2006). *Denken und Problemlösen [Thinking and problem solving].* Göttingen, Germany: Hogrefe.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, *11*, 133-142. doi: 10.1007/s10339-009-0345-0

Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Fontiers in Psychology*(5:739). doi: 10.3389/fpsyg.2014.00739

Funke, J., Fischer, A., & Holt, D. V. (2018). Competencies for complexity: Problem solving in the 21st century. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills* (Vol. 3, p. 41-53). Dordrecht, Netherlands: Springer.

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

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for *g* factor theory and the brief measurement of *g*. *Intelligence*, *52*, 71-79. doi: 10.1016/j.intell.2015.07.006

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, *62*, 138-147. doi: 10.1016/j.intell.2017.04.001

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. doi: 10.1037/a0034716

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 13-23. doi: 10.1016/s0160-2896(97)90011-8

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem solving skills with multiple complex systems. *Thinking & Reasoning*, *21*, 356-382. doi: 10.1016/S0160-2896(97)90011-8

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence*, *41*, 579-596. doi: 10.1016/j.intell.2013.07.012

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* (p. 157-163). Luxembourg, Luxembourg: Office for Official Publications of the European Communities.

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*, 313-317. doi: 10.1027/1015-5759/a000450

Greiff, S., & Neubert, J. C. (2014). On the relation of complex problem solving personality, fluid intelligence, and academic achievement. *Intelligence*, *36*, 37-48. doi: 10.1016/j.lindif.2014.08.003

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36-46. doi: 10.1016/j.chb.2016.02.095

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. doi: 10.1016/j.compedu.2015.10.018

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

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts - something beyond *g*: Concepts, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, *105*, 364-379.

Güss, C. D., Tuason, M. M. T., & Orduña, L. V. (2015). Strategies, tactics, and errors in dynamic decision making in an asian sample. *Journal of Dynamic Decision Making*, *1*(3). doi: 10.11588/jddm.2015.1.13131

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: developmental perspectives on their causal ordering. *Journal of Educational Psychology*, *95*, 123. doi: 10.1037/0022-0663.95.1.124

Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment*, *12*, 1-13. doi: 10.1177/0146621612439620

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., p. 309-341). Charlotte, NC: Age Publishing.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*, 609-618. doi: 10.1198/016214504000000647

Harlen, W. (2005). Trusting teachers' judgment: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*, 245-270. doi: 10.1080/02671520500193744

Hattie, J. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, UK: Routledge. doi: 10.4324/9780203887332

Heine, C., Briedis, K., Didi, H.-J., Haase, K., & Trost, G. (2006). *Bestandsaufnahme von Auswahl- und Eignungsfeststellungsverfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern [Student admission in Germany and selected other countries].* Hannover, Germany: HIS-Kurzinformation A 3/2006.

Helmke, A. (1992). *Selbstvertrauen und schulische Leistungen [Self-confidence and scholastic achievement].* Göttingen, Germany: Hogrefe.

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen [Determinants of scholastic achievement]. In F. E. Weinert (Ed.), *Psychologie des Unterricht und in der Schule* (p. 71-176). Göttingen, Germany: Hogrefe.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*, 297-313. doi: 10.3102/00346543059003297

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi: 10.1080/10705519909540118

Hullemann, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Sciene*, *326*, 1410-1412. doi: 10.1126/science.1177067

Hundertmark, J., Holt, D. V., Fischer, A., Said, N., & Fischer, H. (2015). System structure and cognitive ability as predictors of performance in dynamic system control tasks. *Journal of Dynamic Decision Making*, *1*(5). doi: 10.11588/jddm.2015.1.26416

Huppert, J., Michal-Lomask, S., & Lazarowitz, R. (2002). Computer simulations in the high school: Students' cognitive stages, science process skills, and academic achievement in microbiology. *International Journal of Science Education*, *24*, 803-821. doi: 10.1080/09500690110049150

Ingenkamp, K. (1971). *Die Fragwürdigkeit der Zensurengebung [The dubiousness of school grades].* Weinheim, Germany: Beltz.

Jansen, M., Lüdtke, O., & Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology*, *46*, 116-127. doi: 10.1016/j.cedpsych.2016.05.004

Jansen, M., Schroeders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences. *Learning and Individual Differences*, *30*, 11-21. doi: 10.1016/j.lindif.2013.12.003

Jensen, A. R. (1998a). The *g* factor and the design of education. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment: Theory into practice* (p. 111-131). Mahwah, NJ: Erlbaum.

Jensen, A. R. (1998b). *The g factor: The science of mental ability.* Westport, CT:

Praeger.

Jensen, A. R. (2012). Psychometric g: Definition and substantiation. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (2nd ed., p. 39-53). Mahwah, NJ: Erlbaum.

Jensen, A. R., & Weng, L.-J. (1994). What is a good *g*? *Intelligence*, *18*, 231-258. doi: 10.1016/0160-2896(94)90029-9

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. BIS-Test, Form 4 [Berlin Intelligence-Structure Test. Version 4].* Göttingen, Germany: Hogrefe.

Johnson, W., Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence*, *36*, 81-95.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive *g* and academic achievement *g* one and the same *g*? An exploration of the Woodcock-Johnson and Kaufman tests. *Intelligence*, *40*, 123-138. doi: 10.1016/j.intell.2012.01.009

Kühn, R. (1987). Welche Vorhersagen des Schulerfolgs ermöglichen Intelligenztests? Eine Analyse gebräuchlicher Verfahren [Prediction of academic achievement by means of intelligence measures. An analysis of tests in use]. In R. Horn, K. Ingenkamp, & R. S. Jäger (Eds.), *Tests und Trends 6: Jahrbuch der Pädagogischen Diagnostik* (p. 26-64). München, Germany: Psychologie Verlags Union.

Kind, P. M. (2013). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education*, *97*, 671-694. doi: 10.1002/sce.21070

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

Klauer, K. J., & Marx, E. (2010). Förderung kognitiver Fähigkeiten [Training of cognitive abilities]. In D. H. Rost (Ed.), *Händwörterbuch Pädagogische Psychologie* (4th ed., p. 214-219). Weinheim, Germany: Beltz.

Klauer, K. J., & Sparfeldt, J. R. (2018). Intelligenz und Begabung [Intelligence and ability]. In D. H. Rost, J. R. Sparfeldt, & S. R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie* (5th ed., p. 278-285). Weinheim, Germany: Beltz.

Köller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education*, *32*, 448-470. doi: 10.2307/749801

Künsting, J., Kempf, J., & Wirth, J. (2013). Enhancing scientific discovery learning through metacognitive support. *Contemporary Educational Psychology*, *38*, 349-360. doi: 10.1016/j.cedpsych.2013.07.001

Kretzschmar, A., Neubert, J. C., & Greiff, S. (2014). Komplexes Problemlösen, schulfachliche Kompetenzen und ihre Relation zu Schulnoten [Complex problem solving, school competencies, and their relation to school grades].

Zeitschrift für Pädagogische Psychologie, *28*, 205-215.

Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence*, *54*, 55-69. doi: 10.1016/j.intell.2015.11.004

Kriegbaum, K., Jansen, M., & Spinath, B. (2015). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learning and Individual Differences*, *43*, 140-148. doi: 10.1016/j.lindif.2015.08.026

Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation [Intelligence assessment via computer simulation].* Münster, Germany: Waxmann.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, *33*, 347-368. doi: 10.1016/j.intell.2005.03.002

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*, 63-82. doi: 10.3102/00346543075001063

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, *86*, 148-161. doi: 10.1037/0022-3514.86.1.148

Kwan, J. L. Y., & Chan, W. (2011). Comparing standardized coefficients in structural equation modeling: A model reparameterization approach. *Behavior Research Methods*, *43*, 730-745. doi: 10.3758/s13428-011-0088-6

Lee, T., Cai, L., & MacCallum, R. C. (2012). Power analysis for tests of structural equation models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (p. 181-194). New York, NY: Guilford Press.

LePine, J. A., Colquitt, J. A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive ability, conscientiousness, and openness to experience. *Personnel Psychology*, *53*, 563-569. doi: 10.1111/j.1744-6570.2000.tb00214.x

Leutner, D. (1992). *Adaptive Lehrsysteme: Instruktionspsychologische Grundlagen und experimentelle Analysen [Adaptive instructional systems: Psychological foundations and experimental analyses].* Weinheim, Germany: Psychologie Verlags Union.

Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior*, *18*, 685-697.

Lissmann, U. (2018). Schultests [Scholastic competence tests]. In D. H. Rost, J. R. Sparfeldt, & S. R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie [Concise dictionary of educational psychology]* (5th ed., p. 727-741). Weinheim, Germany: Beltz.

Little, T. D. (2013). *Longitudinal structural equation modeling.* New York, NY: Guilford.

Loehlin, J. C. (1996). The cholesky approach: A cautionary note. *Behavior Genetics*, *26*, 65-69. doi: 10.1007/BF02361160

Lohmann, D. F., & Lakin, J. M. (2011). Intelligence and reasoning. In R. J. Sternberg & S. B. Kaufman (Eds.), *The cambridge handbook of intelligence* (p. 419-441). Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511977244.022

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

Lotz, C., Schneider, R., & Sparfeldt, J. R. (2018). Are intelligence and motivation differentially relevant for scholastic competence tests and grades in mathematics? *Learning and Individual Differences*, *65*, 30-40. doi: 10.1016/j.lindif.2018.03.005

Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex problem solving in educational contexts - still something beyond a "good *g*"? *Intelligence*, *59*, 127-138. doi: 10.1016/j.intell.2016.09.001

MacCallum, R. C. (1986). Specification searches in covariance structure modelling. *Psychological Bulletin*, *100*, 107-120. doi: 10.1037//0033-2909.100.1.107

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490-504. doi: 10.1037//0033-2909.111.3.490

Mackintosh, N. J. (2011). *IQ and Human Intelligence* (2nd ed.). Oxford, UK: Oxford University Press.

Malecki, A. (2016). *Schulen auf einen Blick [Schools at a glance].* Wiesbaden, Germany: Statistisches Bundesamt.

Marsh, H. W., & O'Mara, A. (2009). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology*, *34*, 542-552. doi: 10.1177/0146167207312313

Marsh, H. W., Smith, I. D., Barnes, J., & Butler, S. (1983). Self-concepts: Reliability, stability, dimensionality, validity, and the measurement of change. *Journal of Educational Psychology*, *75*, 772-790. doi: 10.1037//0022-0663.75.5.772

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects model of causal ordering. *Child Development*, *76*, 397-416. doi: 10.1111/j.1467-8624.2005.00853.x

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

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, *82*, 60-70. doi: 10.1037/0022-0663.82.1.60

Mehlhorn, G., & Mehlhorn, H. G. (1981). *Intelligenz. Zur Erforschung und Entwicklung geistiger Fähigkeiten [Intelligence. About the examination and development of cognitive abilities]*. Berlin, DDR: VEB Deutscher Verlag der Wissenschaften.

Ministerium für Bildung und Kultur Saarland. (2013). *Lehrplan Physik Gymnasium Klassenstufen 7 und 8 [curriculum physics gymnasium grade levels 7 and 8].* (Retrieved 2018-03-08, from https://www.saarland.de/dokumente/thema_bildung/LP_Ph_Gym_7_und_8_Mai_2013.pdf)

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of achievement and academic self-concept. *Review of Educational Research*, *79*, 1129-1167. doi: 10.3102/0034654309337522

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J. Morris, & J. L. Amaral (Eds.), *Current topics in children's learning and cognition* (p. 61-82). Rijeka, Croatia: Tech.

Murayama, K., Pekrun, R., Lichtenfeld, S., & vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development*, *84*, 1475–1490. doi: 10.1111/cdev.12036

Naglieri, J. A., & Bornstein, B. T. (2003). Intelligence and achievement: Just how correlated are they? *Journal of Psychoeducational Assessment*, *21*, 244-260. doi: 10.1177/073428290302100302

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., & Urbia, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77-101. doi: 10.1037//0003-066x.51.2.77

Neubert, J. C., Kretzschmar, A., Wüstenberg, S., & Greiff, S. (2015). Extending the assessment of complex problem solving to finite state automata: Embracing heterogeneity. *European Journal of Psychological Assessment*, *31*, 181-194. doi: 10.1027/1015-5759/a000224

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The cambridge handbook of thinking and reasoning* (p. 321-349). Cambridge, NY: University Press. doi: 10.1093/oxfordhb/9780199734689.013.0021

OECD. (2014). *PISA 2012 results: What students know and can do – student* *performance in mathematics, reading, and science.* Paris, France: PISA OECD Publishing.

O'Mara, A. J., Marsh, H. W., Craven, R. G., & Debus, R. L. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist*, *41*, 181-206. doi: 10.1207/s15326985ep4103_4

Pellegrino, J., & Hilton, M. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington, DC: National Academy of Sciences. doi: 10.17226/13398

Powers, D., & Rock, D. (1999). Effects of Coaching on SAT I: Reasoning Test Scores. *Journal of Educational Measurement*, *36*, 93-118. doi: 10.1111/j.1745-3984.1999.tb00549.x

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [On the relation between test intelligence and sucess in problem solving]. *Zeitschrift für Psychologie*, *189*, 79-100.

Putz-Osterloh, W. (1985). Selbstreflektion. Testintelligenz und interindividuelle Unterschiede bei der Bewältigung komplexer Probleme [Self-reflections. Test intelligence and interindividual differences in solving complex problems]. *Sprache & Kognition*, *4*, 203-216.

Putz-Osterloh, W., & Lüer, G. (1981). Über die Vorhersagbarkeit komplexer Problemlöseleistungen durch Ergebnisse in einem Intelligenztest [On the prediction of complex problem solving performance by intelligence test results]. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *28*, 309-334.

Raven, J. C. (1958). *Advanced progressive matrices* (2nd ed.). London, UK: Lewis. doi: 10.1007/springerreference_184521

Raven, J. C. (2000). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology*, *7*, 51-74.

Ree, M. J., & Earles, J. A. (1991). The stability of *g* across different methods of estimation. *Intelligence*, *15*, 271-278. doi: 10.1016/0160-2896(91)90036-d

Reeve, C. L. (2004). Differential ability antecedents of general and specific dimensions of declarative knowledge: More than *g*. *Intelligence*, *32*, 621-652. doi: 10.1016/j.intell.2004.07.006

Reeve, C. L., & Blacksmith, N. (2009). Identifying *g*: A review of current factor analytic practices in the science of mental abilities. *Intelligence*, *37*, 487-494. doi: 10.1016/j.intell.2009.06.002

Rost, D. H. (2013). *Handbuch Intelligenz [Handbook of Intelligence].* Weinheim, Germany: Beltz.

Rost, D. H., & Sparfeldt, J. R. (2002). Facetten des schulischen Selbstkonzepts. Ein Verfahren zur Messung des differentiellen Selbstkonzepts schulischer Leistungen und Fähigkeiten (DISK-Gitter) [Facets of academic self-concept.

Development of a self-concept grid: Psychometric properties and some validity data]. *Diagnostica*, *48*, 130-140. doi: 10.1026//0012-1924.48.3.130

Rost, D. H., Sparfeldt, J. R., & Schilling, S. R. (2007). *DISK-Gitter mit SKSLF-8. Differentielles Schulisches Selbstkonzept mit Skalen zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten (Manual) [DISC grid with SKSLF-8. Differentiated School Self-Concept grid including academic and ability self-concept scales].* Göttingen, Germany: Hogrefe.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118-137. doi: 10.1016/j.intell.2015.09.002

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen: Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen [Intelligence, knowledge, and problem solving: Cognitive prerequisites for successful behavior in computer-simulated problems].* Göttingen, Germany: Hogrefe.

Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen - Perspektiven für eine Kooperation zwischen differentiell-psychometrischer und kognitionspsychologischer Forschung [Intelligence and complex problem solving - Perspectives for a cooperation between differential-psychometric and cognition-psychological research]. *Psychologische Rundschau*, *50*, 220-228. doi: 10.1026//0033-3042.50.4.220

Schaefer, B. A., & McDermott, P. A. (1999). Learning behavior and intelligence as explanations for children's scholastic achievement. *Journal of School Psychology*, *37*, 299-313. doi: 10.1016/s0022-4405(99)00007-2

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*, 102-119. doi: 10.1037//0012-1649.32.1.102

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37-50. doi: 10.1016/j.intell.2014.10.003

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23-74.

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & S. Krapp (Eds.), *The role of interest in learning and development* (p. 183-212). Hillsdale, NJ: Erlbaum.

Schneider, R., Lotz, C., & Sparfeldt, J. R. (2018). Smart, confident, interested: Contributions of intelligence, self-concept, and interest to elementary school achievement. *Learning and Individual Differences*, *62*, 23-35.

Schoenfeld, A. H. (2014). *Mathematical problem solving.* Orlando, FL: Academic Press.

Schoppek, W., & Fischer, A. (2017). Common process demands of two complex dynamic control tasks: Transfer is mediated by comprehensive strategies. *Frontiers in Psychology*, *8:2145*. doi: 10.3389/fpsyg.2017.02145

Schrader, F.-W., & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? [Are teachers' grades influenced by non-achievement-related considerations? An analysis of the determinants of teachers' diagnostic competence]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *22*, 321-324.

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, *24*, 42-52. doi: 10.1016/j.lindif.2012.12.011

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37-63. doi: 10.1016/j.dr.2015.12.001

Skaalvik, E. M., & Valås, H. (1999). Relations among achievement, self-concept, and motivation in mathematics and language arts: A longitudinal study. *Journal of Experimental Education*, *67*, 135-149. doi: 10.1080/00220979909598349

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, A. J., … Latour, T. (2012). The Genetics Lab. Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling*, *54*, 54-72.

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students‘ complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, *41*, 289-305. doi: 10.1016/j.intell.2013.05.002

Sparfeldt, J. R., Buch, S. R., Rost, D. H., & Lehmann, G. (2008). Akkuratesse selbstberichteter Zensuren [Accuracy of self-reported grades]. *Psychologie in Erziehung und Unterricht*, *55*, 68-75.

Sparfeldt, J. R., Rost, D. H., & Schilling, S. R. (2004). Schulfachspezifische Interessen – ökonomisch gemessen [An economical assessment of subject-specific interests in grammar school students]. *Psychologie in Erziehung und Unterricht*, *51*, 213-220.

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., & Thiel, A. (2006). Blocked versus randomized format of questionnaires – a confirmatory multigroup analysis. *Educational and Psychological Measurement*, *66*, 961-974. doi: 10.1177/0013164405285906

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, *15*, 201-293. doi: 10.2307/1412107

Spearman, C. (1923). *The nature of intelligence and the principles of cognition.* London, UK: Macmillan.

Spence, J. T., Pred, R. S., & Helmreich, R. L. (1989). Achievement strivings, scholastic aptitude, and academic perfoprmance: A follow-up to "Impatience versus achievement strivings in the Type A pattern". *Journal of Applied Psychology*, *74*, 176-178. doi: 10.1037//0021-9010.74.1.176

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*, 363-374. doi: 10.1016/j.intell.2005.11.004

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence*, *53*, 92-101. doi: 10.1016/j.intell.2015.09.005

Stankov, L. (1986). Kvashchev's experiment: Can we boost intelligence? *Intelligence*, *10*, 209-230. doi: 10.1016/0160-2896(86)90016-4

Stankov, L. (2012). *g*: A diminutive general. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (2nd ed., p. 19-37). Mahwah, NJ: Erlbaum.

Steinmayr, R., & Meißner, A. (2013). Zur Bedeutung der Intelligenz und des Fähigkeitsselbstkonzeptes bei der Vorhersage von Leistungstests und Noten in Mathematik. [The importance of intelligence and ability self-concept for the prediction of standardized achievement tests and grades in mathematics]. *Zeitschrift für Pädagogische Psychologie*, *27*, 273-282. doi: 10.1024/1010-0652/a000113

Steinmayr, R., Meißner, A., Weidinger, A. F., & Wirthwein, L. (2014). Academic achievement. In L. H. Meyer (Ed.), *Oxford bibliographies online: Education.* New York, NY: Oxford University Press. doi: 10.1093/obo/9780199756810-0108

Steinmayr, R., Sauer, J., & Gamsjäger, E. (2018). Prognose von Schulerfolg [Predicting educational success]. In D. H. Rost, J. R. Sparfeldt, & S. R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie [Concise dictionary of educational psychology]* (5th ed., p. 653-665). Weinheim, Germany: Beltz.

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, *19*, 80-90. doi: 10.1016/j.lindif.2008.05.004

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, *47*, 1-41. doi: 10.1353/mpq.2001.0005

Strohschneider, S., & Güss, C. D. (1999). The fate of the Moros: A cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology*, *34*, 235-252. doi: 10.1080/002075999399873

Tajudin, N. M., & Chinnappan, M. (2015). Exploring relationship between scientific reasoning skills and mathematics problem solving. In M. Marshman, V. Geiger, & A. Bennison (Eds.), *Mathematics education in the margins* (p. 603-610). Sunshine Coast, Australia: MERGA.

Tent, L., Fingerhut, W., & Langfeldt, H.-P.-. (1976). *Quellen des Lehrerurteils [Sources of teacher assessments].* Weinheim, Germany: Beltz.

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer-based assessment. *Computers & Education*, *56*, 1032-1044. doi: 10.1016/j.compedu.2010.11.017

Thorndike, E. L. (1922). Practice effects in intelligence tests. *Journal of Experimental Psychology*, *5*, 101-107. doi: 10.1037/h0074568

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy-value theory: A latent interaction modeling study. *Journal of Educational Psychology*, *104*, 763-777. doi: 10.1037/a0027470

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*, 1-10. doi: 10.2307/1129583

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefes and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*, 111-133. doi: 10.1207/s15326985ep3902_3

Valerius, S., & Sparfeldt, J. R. (2014). Consistent *g*- as well as consistent verbal-, numerical- and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries. *Intelligence*, *44*, 120-133. doi: 10.1016/j.intell.2014.04.003

Valerius, S., & Sparfeldt, J. R. (2015). Zusammenhänge allgemeiner und spezifischer Intelligenzfaktoren mit allgemeinen und spezifischen Schulleistungen im Nested-Factor-Modell [Relations of general and specific intelligence factors with general and specific achievement factors in a nested-factor-model]. *Zeitschrift für Pädagogische Psychologie*, *29*, 101-108. doi: 10.1024/1010-0652/a000151.

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, *43*, 381-400. doi: 10.1007/s11251-015-9344-y

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences*, *29*, 123-130. doi: 10.1016/j.lindif.2013.01.003

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*, 89-109. doi: 10.1016/j.learninstruc.2003.10.004

Vieluf, U., Ivanov, S., & Nikolova, R. (2011). *Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe [Competencies and attitudes of school students in schools of Hamburg at the end of the first stage of secondary education and at the beginning of the second stage of secondary education].*

Münster, Germany: Waxmann.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, *20*, 75-100. doi: 10.1207/s15516709cog2001_3

Weber, H. S., Lu, L., Shi, J., & Spinath, F. M. (2013). The roles of cognitive and motivational predictors in explaining school achievement in elementary school. *Learning and Individual Differences*, *25*, 85-92. doi: 10.1016/j.lindif.2013.03.008

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 – Revision (CFT 20-R) [Culture Fair Intelligence Test 20-R–Scale 2].* Göttingen, Germany: Hogrefe.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (p. 209-231). New York, NY: Guilford Press.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68-81. doi: 10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (p. 173-195). San Diego, CA: Academic Press. doi: 10.1016/b978-012750053-9/50006-1

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1-37. doi: 10.1111/j.1745-3984.2002.tb01133.x

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In R. B. Catell & J. R. Nesselroade (Eds.), *Handbook of multivariate experimental psychology* (p. 505-560). New York, NY: Plenum Press.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*, 913-934. doi: 10.1177/0013164413495237

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

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. doi: 10.1016/j.lindif.2013.10.006

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology Knowledge and Learning*, *19*, 127-146. doi: 10.1007/s10758-014-9222-8

Zaunbauer, A. C. M., Retelsdorf, J., & Möller, J. (2009). Die Vorhersage von Englischleistungen am Anfang der Sekundarstufe [Prediction of english achievement in early secondary school]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *41*, 153-164. doi: 10.1026/0049-8637.41.3.153

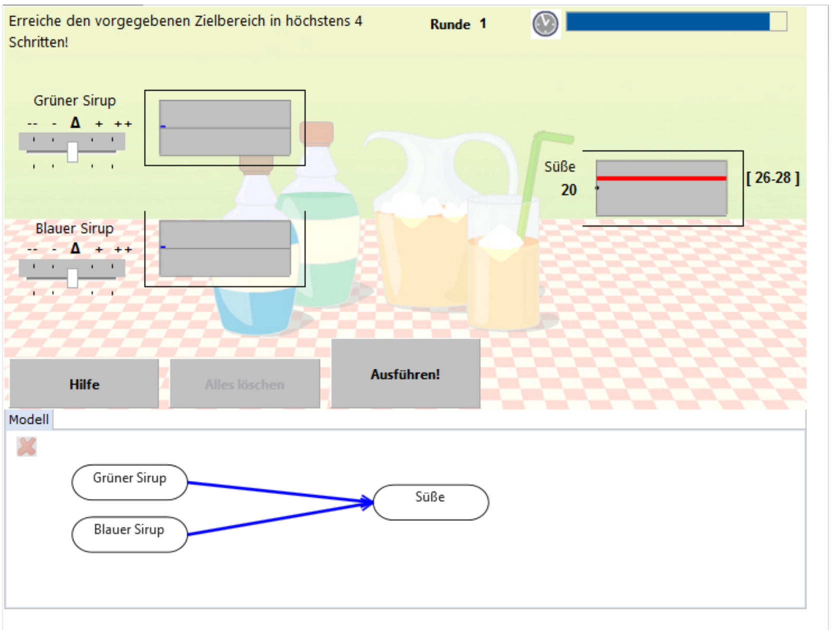
Zimmerman, C., & Croker, S. (2013). Learning science through inquiry. In G. Feist & M. Gorman (Eds.), *Handbook of the psychology of science* (p. 49-70). New York, NY: Springer.

Zuffianò, A., Alessandri, G., Gerbino, M., Kanacri, B. P. L., Giunta, L. D., Milioni, M., & Caprara, G. V. (2013). Academic achievement: The unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. *Learning and Individual Differences*, *23*, 158-162. doi: 10.1016/j.lindif.2012.07.010

Appendix

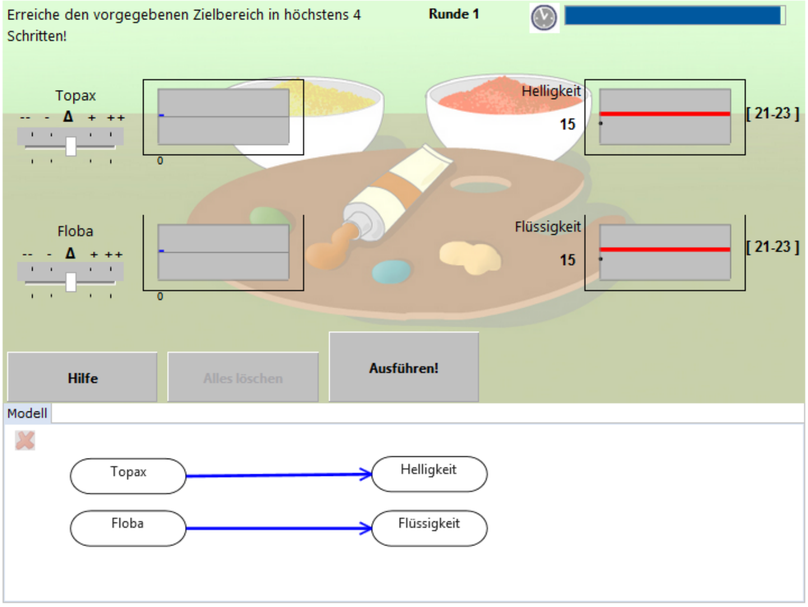
A MicroDYN tasks

Screenshots of the nine MicroDYN tasks that were used in this dissertation project. The to be explored models within the problem space and the number and type of the minimally required exploration steps are displayed.



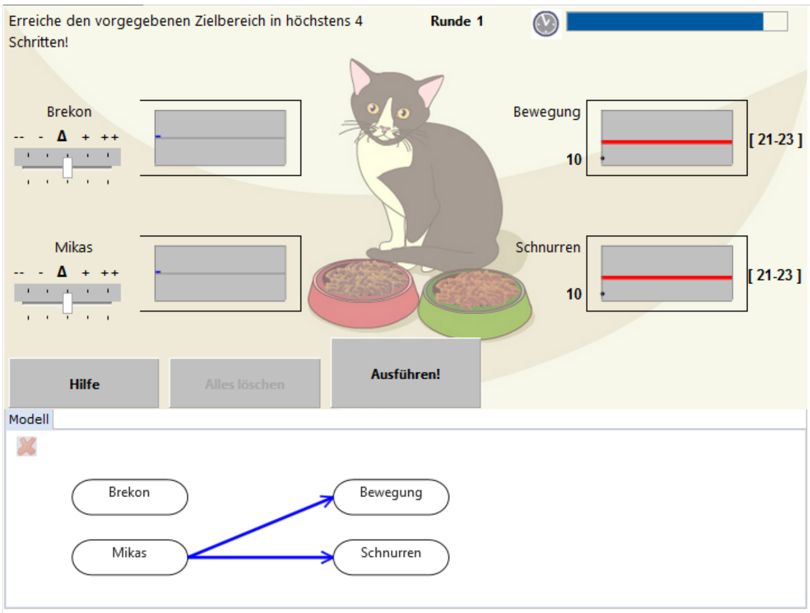
Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.1: Screenshot of the first MicroDYN task (Lemonade); Minimally required exploration steps: 2 × VOTAT (once for each input variable).



Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.2: Screenshot of the second MicroDYN task (Drawing); Minimally required exploration steps: 2 × VOTAT (once for each input variable).



Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.3: Screenshot of the third MicroDYN task (Cat); Minimally required exploration steps: 2 × VOTAT (once for each input variable).

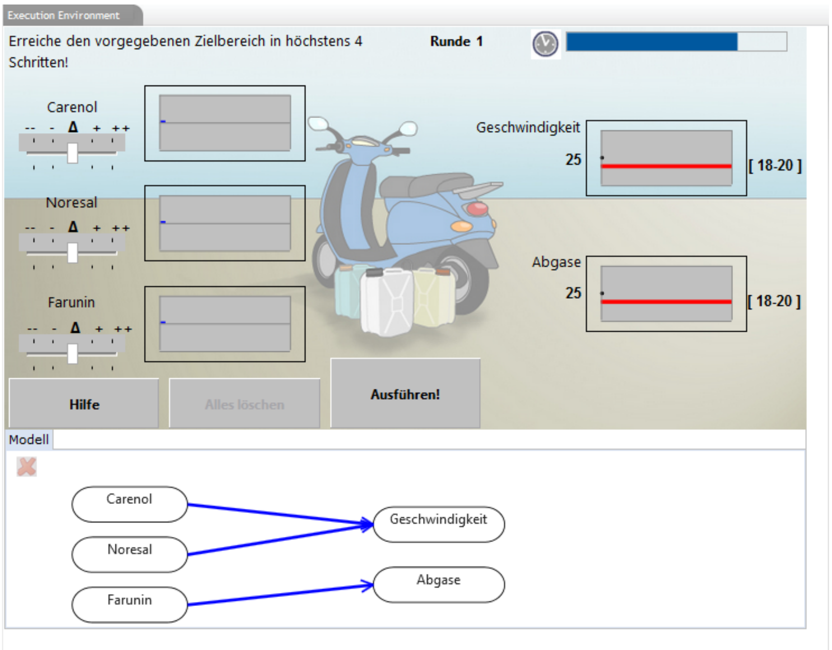


Figure A.4: Screenshot of the fourth MicroDYN task (Moped); Minimally required exploration steps: 3 × VOTAT (once for each input variable).

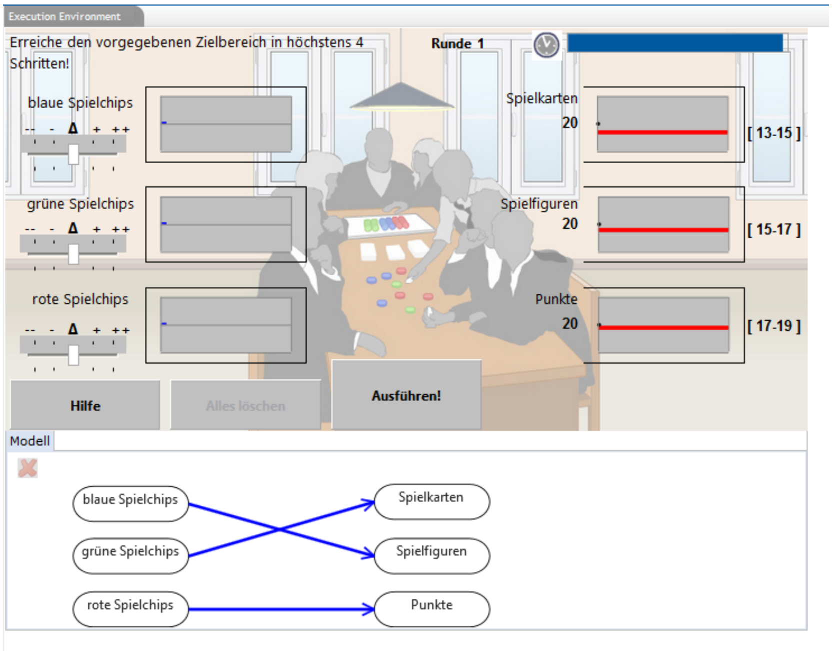
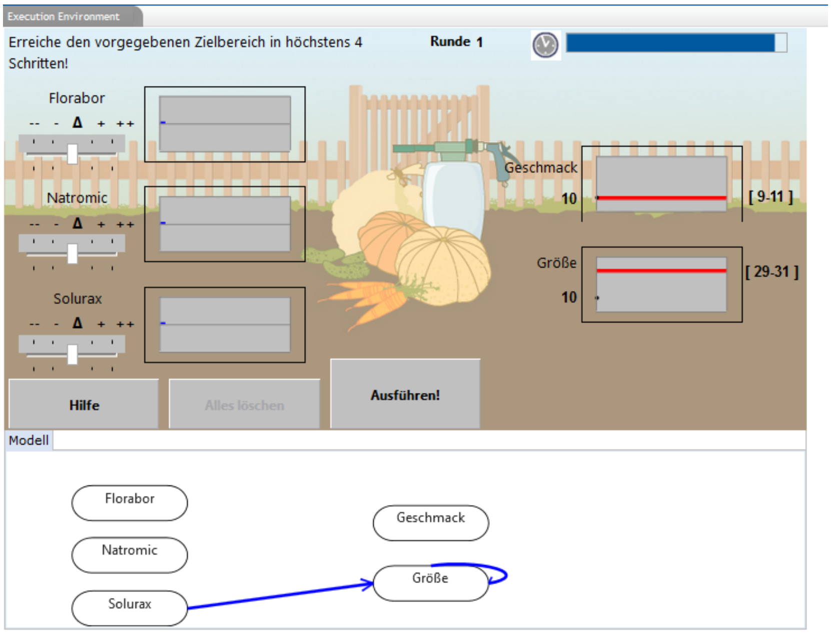
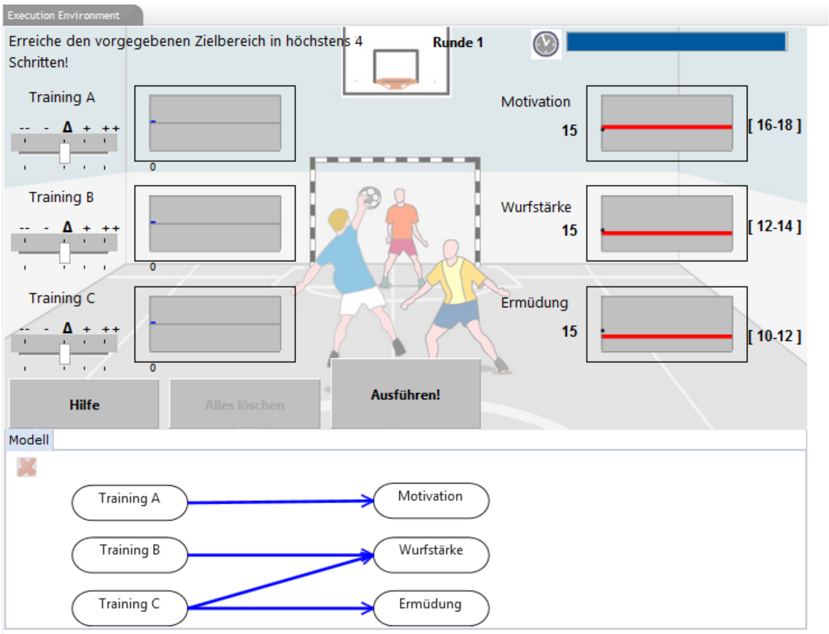


Figure A.5: Screenshot of the fifth MicroDYN task (Gaming); Minimally required exploration steps: 3 × VOTAT (once for each input variable).



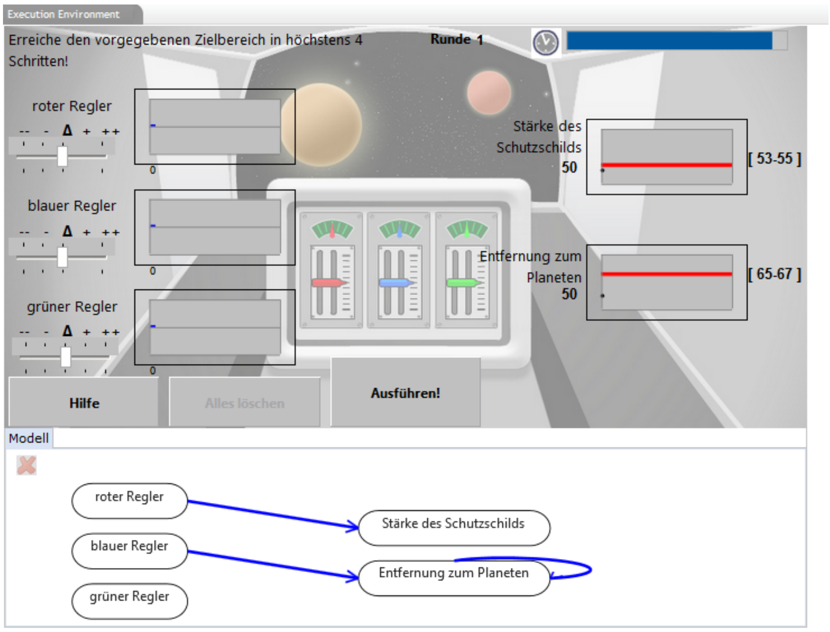
Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.6: Screenshot of the sixth MicroDYN task (Gardening); Minimally required exploration steps: 3 × VOTAT (once for each input variable); 1 × NOTAT.



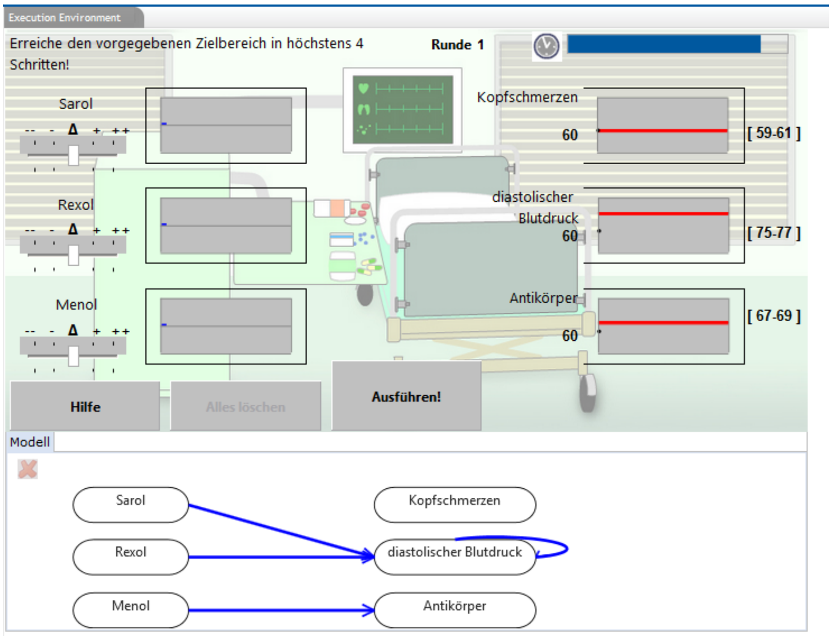
Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.7: Screenshot of the seventh MicroDYN task (Handball); Minimally required exploration steps: 3 × VOTAT (once for each input variable).



Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.8: Screenshot of the eighth MicroDYN task (Space); Minimally required exploration steps: 3 × VOTAT (once for each input variable); 1 × NOTAT.



Erreiche den vorgegebenen Zielbereich in höchstens 4 Schritten!

Figure A.9: Screenshot of the ninth MicroDYN task (Medical Aid); Minimally required exploration steps: 3 × VOTAT (once for each input variable); 1 × NOTAT.

B Exemplary M*plus* Syntaxes

Exemplary M*plus* syntaxes that were used for conducting the analyses in the three articles of this dissertation project are provided below.

B.1 Exemplary M*plus* syntax of the regression-based model in mathematics (Lotz, Sparfeldt, & Greiff, 2016)

```
title:
Regression-based Model of Intelligence and CPS for conducting the path comparisons;


data: file = CPS_Math.dat;

variable: names = KL_ID OG XG BD ST WA ZN AN SC CH BM
Mod_m Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Gar Mod_Han Mod_Spa Mod_Aid
Forc_m for_Lem for_Dra for_Cat for_Mop for_Gam for_Gar for_Han for_Spac for_Aid
M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
M11r M12r M13r M14r M15r M16r M17r M18r M19r M20r
21r M22r M23r M24r M25r M26r M27r M28r M29r M30r GradeM;

!CPS items and items of the mathematics competence test were dichotomous
categorical = Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Han
              for_Lem for_Dra for_Cat for_Mop for_Gam for_Han
              for_Spac for_Aid

              M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
              M11r M12r M13r M14r M15r M16r M17r M18r M19r
              M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;

usevar = !class id
       KL_ID

     !10 Intelligence subtests
          WA ZN AN SC OG XG BD ST CH TM

     !CPS items
      Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Han
            for_Lem for_Dra for_Cat for_Mop for_Gam
            for_Han for_Spac for_Aid
      !without item for_Gar because no student solved this item corretly

         !30 mathematics competence test items
           M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
           M11r M12r M13r M14r M15r M16r M17r M18r M19r
           M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r

         !Reversly scored grade in mathematics
           GradeM;

missing = all(-99);      !Missing values are identified by -99
cluster = KL_ID;       !Observations are clustered within classes
analysis: type = complex; !Analysis takes clustering within classes into account

model:

!!!Measurement models!!!
!CPS

  CPS_Mod by Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Han;
      !without Mod_Gar Mod_Spa Mod_Aid because of defect item builder

  CPS_For by for_Lem for_Dra for_Cat for_Mop for_Gam for_Han
         for_Spac for_Aid;


  CPS by CPS_Mod@1 CPS_for@1; !Loading of the first-order CPS factors were
                 !constrained to be equal to avoid under-identification

!Intelligence

  BIS_v by TM Wa ST;       !defining the verbal content factor
  BIS_n by ZN XG SC;       !defining the numerical content factor
  BIS_f by CH OG AN BD;    !defining the figural content factor

  G by BIS_v* BIS_n BIS_f;  !defining the second-order g factor
```

```

!Variances of the latent intelligence and CPS factors were constrained to be equal
  BIS_G (1);
  KPL (1);


!Mathematics competence test
  Math by M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
      M11r M12r M13r M14r M15r M16r M17r M18r M19r
      M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;

!!!Structural model!!!

CPS with G;          !CPS correlates with intelligence
GradeM with Math;    !Grade correlates with athematics competence test

GradeM on CPS G;     !Grade is regressed on CPS and intelligence
Math on CPS G;       !Mathematics competence test is regressed on CPS and intelligence

!File Math_Basic.dat serves as the comparision model for later path comparisons
Savedata: difftest is Math_Basic.dat;

output: sampstat standardized tech4 stdyx modindices (3.84);
```

Listing 1: M*plus* syntax of the regression-based model in mathematics (Lotz, Sparfeldt, & Greiff, 2016)

B.2 Exemplary *Mplus* syntax of the CPS-residual model in mathematics (Lotz, Sparfeldt, & Greiff, 2016)

```
title:
CPS Residual Model, Intelligence and the increment of CPS predict
the mathematics competence test and grades;

data: file = CPS_Math.dat;

variable: names = KL_ID OG XG BD ST WA ZN AN SC CH BM
Mod_m Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Gar Mod_Han Mod_Spa Mod_Aid
Forc_m for_Lem for_Dra for_Cat for_Mop for_Gam for_Gar for_Han for_Spac for_Aid
M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
M11r M12r M13r M14r M15r M16r M17r M18r M19r M20r
21r M22r M23r M24r M25r M26r M27r M28r M29r M30r GradeM;

!CPS items and items of the mathematics competence test were dichotomous
categorical = Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Han
              for_Lem for_Dra for_Cat for_Mop for_Gam for_Han
              for_Spac for_Aid

              M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
              M11r M12r M13r M14r M15r M16r M17r M18r M19r
              M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;

usevar = !class id
       KL_ID

     !10 Intelligence subtests
          WA ZN AN SC OG XG BD ST CH TM

     !CPS items
      Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Han
            for_Lem for_Dra for_Cat for_Mop for_Gam
            for_Han for_Spac for_Aid
      !without item for_Gar because no student solved this item corretly

         !30 mathematics competence test items
           M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
           M11r M12r M13r M14r M15r M16r M17r M18r M19r
           M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r

         !Reversly scored grade in mathematics
           GradeM;

missing = all(-99);     !Missing values are identified by -99
cluster = KL_ID;        !Observations are clustered within classes
analysis: type = complex; !Analysis takes clustering within classes into account

model:

!!!Measurement models!!!
!CPS

  CPS_Mod by Mod_Lem Mod_Dra Mod_Cat Mod_Mop Mod_Gam Mod_Han;
      !without Mod_Gar Mod_Spa Mod_Aid because of defect item builder

  CPS_For by for_Lem for_Dra for_Cat for_Mop for_Gam for_Han
         for_Spac for_Aid;


  CPS by CPS_Mod@1 CPS_for@1; !Loading of the first-order CPS factors were
                !constrained to be equal to avoid under-identification

!Intelligence

  BIS_v by TM Wa ST;        !defining the verbal content factor
  BIS_n by ZN XG SC;        !defining the numerical content factor
  BIS_f by CH OG AN BD;     !defining the figural content factor

  G by BIS_v BIS_n BIS_f;    !defining the second-order g factor



!Mathematics competence test
  Math by M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
      M11r M12r M13r M14r M15r M16r M17r M18r M19r
      M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;

!!!Structural model!!!
```

```

CPSres by CPS;     !CPS residual factor is indicated by the CPS factor
CPS@0;          !Variance of the CPS factor is fixed to zero
CPSres with G@0;  !Rediual factor of CPS is not correlated with intelligence
CPS on G;       !CPS factor is regressed on intelligence

Note_Mr on CPSres G;!Grade is regressed on CPS Residual and intelligence
Mathe on CPSLres G; !Mathematics competence test is regressed on CPS Residual and intelligence




output: sampstat standardized tech4 stdyx modindices (3.84);
```

Listing 2: M*plus* syntax of the CPS-residual model in mathematics (Lotz, Sparfeldt, & Greiff, 2016)

B.3 Exemplary M*plus* syntax of the augmented VOTAT LGCM (Lotz, Scherer, Greiff, & Sparfeldt, 2017)

```
title:
Discontinuouse Latent Growth Curve Model for 9 VOTAT items (relative frequency),
tasks type change after task 5,
phases before and after the task type chanfge are represented by separate intercept ans slope factors
latent intercept and slope factors correlate with intelligence;

data: file = CPS_Processdata.dat;

variable: names = KL_ID WA ZN AN SC CH TM OG XG BD ST
          Vot1 Vot2 Vot3 Vot4 Vot5 Vot6 Vot7 Vot8 Vot9;

usevar = !class id
       KL_ID
     !10 Intelligence subtests of the BIS−4 battery
           WA ZN AN SC OG XG BD ST CH TM
         !relative VOTAT frequency items
           Vot1 Vot2 Vot3 Vot4 Vot5 Vot6 Vot7 Vot8 Vot9;


missing = all(−99);        !Missing values are identified by −99
cluster = KL_ID;         !Observations are clustered within classes
analysis: type = complex;   !Analysis takes clustering within classes into account

model:

!!!Discontinuouse Laten Growth Curve Model!!!
!Defining the intercept and slope factors before the task type change
!Item 5 (directly before task type change) is reference item, i.e. fixed tu zero
  Int_15 Slope_15 | Vot1@−4 Vot2@−3 Vot3@−2 Vot4@−1 Vot5@0;

!Defining the intercept and slope factors after the task type change
!Item 6 (directly after task type change) is reference item, i.e. fixed tu zero
  Int_69 Slope_69 | Vot6@0 Vot7@1 Vot8@2 Vot9@3;

!!!Intelligence part of the model!!!
!Measurement model
  BIS_v by ST TM WA;       !defining the verbal content factor
  BIS_n by ZN XG SC;       !defining the numerical content factor
  BIS_f by CH AN OG BD;    !defining the figural content factor

  BIS_G by BIS_n BIS_v BIS_f; !defining the second−order g factor

!modeling the residual correlations among the items of the reasoning operation facet
  TM with WA ZN SC CH AN;
  WA with ZN SC CH AN;
  ZN with SC CH AN;
  SC with CH AN;
  CH with AN;

!modeling the residual correlations among the items of the speed operation facet
  ST with OG;

!modeling the residual correlations among the items of the memory operation facet
  XG with BD;


!!!Structural model!!!
!Intelligence correlates with the intercept and slope factors

  BIS_G with Int_15 Slope_15 Int_69 Slope_69;


output: sampstat standardized tech4 stdyx modindices (3.84);
```

Listing 3: M*plus* syntax of the augmented VOTAT LGCM (Lotz, Scherer, Greiff, & Sparfeldt, 2017)

B.4 Exemplary M*plus* syntax of the reparameterized Intelligence Self-concept Model (Lotz, Schneider, & Sparfeldt, 2018)

```
title:
reparameterized Intelligence Self-concept Model for cunducting the comparisons
of the standardized paths;


data: file = g_SC_Inter_math.dat;

variable: names = KL_ID OG XG BD ST WA ZN AN SC CH TM
SCM1 SCM2 SCM3 SCM4 SCM5 SCM6 SCM7 SCM8
M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r M11r M12r M13r M14r M15r
M16r M17r M18r M19r M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r
GradeM;

!Items of the mathematics competence test were dichotomous
categorical = M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
              M11r M12r M13r M14r M15r M16r M17r M18r M19r
              M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;

usevar = !class id
       KL_ID

     !10 Intelligence subtests
          WA ZN AN SC OG XG BD ST CH TM

         !8 Mathematics self-concept items
           SCM1 SCM2 SCM3 SCM4 SCM5 SCM6 SCM7 SCM8

         !30 mathematics competence test items
           M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
           M11r M12r M13r M14r M15r M16r M17r M18r M19r
           M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r

         !Reversly scored grade in mathematics
           GradeM;

missing = all(-99);     !Missing values are identified by -99
cluster = KL_ID;       !Observations are clustered within classes
analysis: type = complex; !Analysis takes clustering within classes into account

model:

!!!Measurement models!!!
!Intelligence

  BIS_v by TM Wa ST;       !defining the verbal content factor
  BIS_n by ZN XG SC;       !defining the numerical content factor
  BIS_f by CH OG AN BD;    !defining the figural content factor

  G by BIS_v BIS_n BIS_f;    !defining the second-order g factor

!Self-concept
  SC by SCM1 SCM2 SCM3 SCM4 SCM5 SCM6 SCM7 SCM8;

!Mathematics competence test
  Math by M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
       M11r M12r M13r M14r M15r M16r M17r M18r M19r M20r
       M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;

!Automatic correlations are fixed to zero
  BIS_v with Math@0 GradeM@0 SC@0 G@0 BIS_n@0 BIS_f@0;
  BIS_n with Math@0 GradeM@0 SC@0 G@0 BIS_f@0;
  BIS_f with Math@0 GradeM@0 SC@0 G@0;

!!!Reparameterization!!!

!Fixing variances to Zero
  G@0;
  SC@0;
  Math@0;
  Math_PH1@0;
  Math_PH2@0;
  GradeM@0;
  GrM_PH1@0;
  GrM_PH2@0;


!Regressing the original variables (G, SC, Math, GradeM) on their
```

```
phantom variables (G_PH1, SC_PH1, Math_PH1, GrM_PH1)

!Intelligence
  G_PH1 by G*;
  G_PH1@1;     !Fixing variance of the Phantom variable (G_PH1) to 1

!Self-concept
  SC_PH1 by SC*;
  SC_PH1@1;   !Fixing variance of the Phantom variable (SC_PH1) to 1

!Mathematics competence test
  Math_PH1 by Math*;

!Mathematics Grade
G rM_PH1 by GradeM*;

!Variances of the dependent variables phantom variables are not fixed to 1

!Regressing the dependent variables phantom variables (Math_PH1, GrM_PH1) on their
disturbance terms (Math_PH2, GrM_PH2);
!paths are fixed to 1
  Math_PH2 by Math_PH1@1;
  GrM_PH2 by GrM_PH1 @1;


!Regressing the disturbance terms (Math_PH2, GrM_PH2) on their phantom variables (F999, F998);
!Paths are fixed to 1
!Variance of the phantom variables (F999, F998) are fixed to 1
  F999 by Math_PH2@1;
  F999@1;

  F998 by GrM_PH2@1;
  F998@1;


!Structural model of the phantom variables (G_PH1, SC_PH1, Math_PH1, !GrM_PH1)
  G_PH1 with SC_PH1 (c1);

  Math_PH1 on G_PH1 (c12);
  Math_PH1 on SC_PH1 (c13);

  GrM_PH1 on G_PH1 (c15);
  GrM_PH1 on SC_PH1 (c16);

!Structural model of the image variables (G_PH2, SC_PH2, Math_PH2, GrM_PH2)
  G_PH2 with SC_PH2 (c5);

  G_PH2 by Math_PH2* (c12);
  SC_PH2 by Math_PH2* (c13);

  G_PH2 by GrM_PH2* (c15);
  SC_PH2 by GrM_PH2* (c16);

! e.g. (c12) --> Paths of the image variables structural model are
constraint to be equal to the paths of the phantom variables structural model

!Variance of the independents variables image variables are fixed to -1
  SC_PH2@-1;
  G_PH2@-1;

!automatic correlations are fixed to zero
  G_PH2 with G_PH1@0 SC_PH1@0  F999@0 F998@0;
  SC_PH2 with SC_PH1@0 G_PH1@0 F999@0 F998@0;
  F999 with SC_PH1@0 G_PH1@0;
  F998 with SC_PH1@0 G_PH1@0;

!Covariance of the independent variables image variables is constrained to
be equal to the negative covariance of the independent variables phantom variables
Model constraint:
  c1=-c5;


output: sampstat standardized tech4 stdyx modindices (3.84);
```

Listing 4: M*plus* syntax of the reparameterized Intelligence Self-concept Model (Lotz, Schneider, & Sparfeldt, 2018)

B.5 Exemplary *Mplus* syntax of the Cholesky factoring of the Self-concept Increment Model (Lotz, Schneider, & Sparfeldt, 2018)

```
title:
Exemplarily Mplus syntax for the cholesky factoring of the Self-concept Increment Model

data: file = g_SC_Inter_math.dat;

variable: names = KL_ID OG XG BD ST WA ZN AN SC CH TM SCM1 SCM2 SCM3 SCM4 SCM5 SCM6 SCM7 SCM8 INTRM1 INTRM2
     INTRM3 INTRM4 M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r M11r M12r M13r M14r M15r M16r M17r M18r M19r
     M20r M21r
M22r M23r M24r M25r M26r M27r M28r M29r M30r GradeM;

!Items of the mathematics competence test were dichotomous
categorical = M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
              M11r M12r M13r M14r M15r M16r M17r M18r M19r
              M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;


usevar = !class id
       KL_ID

     !10 Intelligence subtests
          WA ZN AN SC OG XG BD ST CH TM

         !8 Mathematics self-concept items
           SCM1 SCM2 SCM3 SCM4 SCM5 SCM6 SCM7 SCM8
         !Mathematics interest
           INTRM1 INTRM2 INTRM3 INTRM4

         !30 mathematics competence test items
           M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
           M11r M12r M13r M14r M15r M16r M17r M18r M19r
           M20r M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r

         !Reversly scored grade in mathematics
           GradeM;

missing = all(-99);      !Missing values are identified by -99
cluster = KL_ID;       !Observations are clustered within classes
analysis: type = complex; !Analysis takes clustering within classes into account

model:

!!!Measurement models!!!
!Intelligence

  BIS_v by TM Wa ST;        !defining the verbal content factor
  BIS_n by ZN XG SC;        !defining the numerical content factor
  BIS_f by CH OG AN BD;     !defining the figural content factor

  G by BIS_v BIS_n BIS_f;     !defining the second-order g factor

!Self-concept
  SC by SCM1 SCM2 SCM3 SCM4 SCM5 SCM6 SCM7 SCM8;

!Interest
  INTR by INTRM9 INTRM10 INTRM11 INTRM12;

!Mathematics competence test
  Math by M1r M2r M3r M4r M5r M6r M7r M8r M9r M10r
      M11r M12r M13r M14r M15r M16r M17r M18r M19r M20r
      M21r M22r M23r M24r M25r M26r M27r M28r M29r M30r;





!Automatic correlations are fixed to zero
  BIS_v with Math@0 GradeM@0 SC@0 G@0 BIS_n@0 BIS_f@0;
  BIS_n with Math@0 GradeM@0 SC@0 G@0 BIS_f@0;
  BIS_f with Math@0 GradeM@0 SC@0 G@0;

!!!Structural model!!!

!Defining the phantom variables
  G_PH by G* SC INTR;       !Intelligence was assigned first priority
  INTR_PH by INTR* SC;      !Interest was assigned second priority
  SC_PH by SC*;         !Self-concept was assigned last priority

```

```
!Variance of the phantom variables is fixed to 1
  G_PH@1;
  SC_PH@1;
  INTR_PH@1;

!Variance of the original variables is fixed to zero
  G@0;
  SC@0;
  INTR@0;

!automatic correlations of the phantom variables are fixed to zero
  G_PH with SC_PH@0 INTR_PH@0;
  SC_PH with INTR_PH@0;

!Regression model with phantom variables as predictors
  Math on G_PH INTR_PH SC_PH;
  GradeM on G_PH INTR_PH SC_PH;


output: sampstat standardized tech4 stdyx modindices (3.84);
```

Listing 5: M*plus* syntax of the Cholesky factoring of the Self-concept Increment Model (Lotz, Schneider, & Sparfeldt, 2018)

C Publications

Appendix C1: Publication 1

Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex problem solving in educational contexts — Still something beyond a ”good *g*”? *Intelligence, 59,* 127–138. doi: 10.1016/j.intell.2016.09.001

Appendix C2: Publication 2

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

Appendix C3: Publication 3

Lotz, C., Schneider, R., & Sparfeldt, J. R. (2018). Are intelligence and motivation differentially relevant for scholastic competence tests and grades in mathematics? *Learning and Individual Differences, 65*, 30–40. doi: 10.1016/j.lindif.2018.03.005