

DISS. ETH NO. 22896

***THE INTERPLAY BETWEEN GENDER,
UNDERACHIEVEMENT, AND CONCEPTUAL
INSTRUCTION IN PHYSICS***

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

SARAH ISABELLE HOFER

Diplom-Psychologin Univ., Ludwig-Maximilians-Universität München

born on 30.07.1985
citizen of Germany

accepted on the recommendation of

Prof. Dr. Elsbeth Stern
Prof. Dr. Andreas Vaterlaus
Prof. Dr. Oliver Lüdtke

2015

Abstract

In physics classrooms, many students fail to understand the relevant concepts. Even highly intelligent students frequently struggle with acquiring conceptual understanding in physics. This struggle seems to be especially pronounced among female students. The students' difficulties to understand physics concepts have been extensively shown in the field of Newtonian mechanics, where even students who are able to apply the formulae correctly often hold severe misconceptions. These findings result from the complex interplay between characteristics of the student, the assessment, and the instruction in the physics classroom. The four papers of this thesis investigate this interplay, focusing on gender differences, underachievement, and conceptual instruction in secondary school physics. In conjunction, the four papers demonstrate what can be done to help the students, both female and male, to invest their intellectual potential in understanding physics.

The first paper concentrates on the identification of underachievers. Underachievers are students whose physics grades considerably lag behind their high intelligence test performance. The paper applies a new approach to identify underachievers by means of latent profile analysis. The latent profile analysis was implemented in a sample of $N = 316$ Swiss secondary school students. The resulting profiles proved to be theoretically meaningful and methodically valid. A profile of physics underachievers with high intellectual potential but below average physics grades was detected only among the girls. These female underachievers were further characterized by low interest and self-concept in physics.

The second paper addresses the contribution of physics teachers' assessment practice to explaining the high portion of female underachievers. Using an online-survey, this paper examined whether physics teachers' assessment of student performance varies systematically as a function of the students' gender. A sample of $N = 780$ physics teachers from Switzerland, Austria, and Germany participated. The teachers graded a fictive female or male student's answer to a physics test question. The identical answer was neither completely wrong nor absolutely correct, leaving room for interpretation. In addition to the country of origin, the teachers' gender and length of teaching experience were inquired to allow comprehensive analyses. The results indicated a gender bias against girls for all Swiss and Austrian teachers, as well as for German female teachers in the first decade of their career. The partially correct answer received worse evaluations, when the answer was randomly

assigned to originate from a female student. Yet, the bias disappeared with increasing teaching experience. German male teachers showed no gender bias effects at all. The findings of this study suggest that, in assessment situations that leave room for interpretation, male students more often profit from a benefit of the doubt than female students. More generally, the deviations in the teachers' assessment of a student's answer point out that fair assessment of student performance is a difficult process that requires particular attention.

The third paper hence elaborates on the process of assessment and focusses on the assessment of conceptual understanding. Common assessment in physics classrooms, by contrast, focusses on quantitative problem solving. The paper describes the development and evaluation of the "basic Mechanics Concept Test" (bMCT). The bMCT is a multiple choice test that assesses fundamental conceptual understanding in Newtonian mechanics adapted to the content taught to secondary school students. The development of the test was based on a sample of $N = 239$ students. Subsequently, an evaluation study with $N = 141$ students confirmed that the bMCT satisfies the Rasch model. The analyses show that the bMCT validly assesses conceptual understanding and conceptual development regarding Newtonian mechanics.

The fourth paper combines the different aspects covered in the first three papers. In addition, the comprehensive intervention study that is described in the fourth paper took characteristics of the instruction into account. With a sample of $N = 172$ students, it examined cognitively activating conceptual instruction that specifically incorporated elements that foster conceptual understanding. The study investigated the potential of cognitively activating conceptual instruction to address deficient physics knowledge in general and tackle underachievement in particular. Four classes that had received 18 lessons of cognitively activating conceptual instruction were compared with four classes that had received 18 lessons of conventional instruction on Newtonian mechanics. Four physics teachers taught one class according to a cognitively activating teaching manual and one class as always (i.e., conventional instruction). Quantitative problem solving and conceptual understanding (using the bMCT) were assessed together with motivational variables across three points of measurement (pre, post, and follow-up). Students benefitted from cognitively activating conceptual instruction in terms of both performance measures, but not in terms of motivation. Latent profile analyses revealed that female underachievers profited considerably from cognitively activating conceptual instruction with respect to their follow-up conceptual

understanding, although this performance boost was not reflected immediately in the posttest and the grades.

Each of the four studies has implications for physics classrooms and educational research in its own right. Taken together, they provide important insights into the interplay between gender, underachievement, and conceptual instruction in secondary school physics classrooms that go beyond the scope of each individual paper. In a synopsis of the findings, three overarching issues are discussed: factors that may contribute to underachievement, means to tackle the gender gap in physics, as well as the general design of physics lessons. The recommendations for future research and educational practice that are derived in the course of this final discussion may help to enable more students, female and male, to invest their intellectual potential in understanding physics.

Zusammenfassung

Im Physikunterricht scheitern selbst überdurchschnittlich intelligente Schülerinnen und Schüler oft am Verständnis der relevanten Konzepte. Bei Schülerinnen scheint dies besonders ausgeprägt zu sein. Verständnisschwierigkeiten wurden vor allem auf dem Gebiet der Newtonschen Mechanik vielfach nachgewiesen, wo auch Schülerinnen und Schüler, die mit den Formeln umgehen können, profunden konzeptuellen Fehlvorstellungen unterliegen. Die Ursachen hierfür sind vielfältig und das Resultat eines komplexen Zusammenspiels von Merkmalen der Lernenden, der Leistungsbewertung und der Instruktion im Physikunterricht. Die vier Artikel dieser Arbeit befassen sich mit eben diesem Zusammenspiel, wobei ein Schwerpunkt auf Geschlechtsunterschieden, Underachievement und konzeptbasierter Instruktion im gymnasialen Physikunterricht liegt. Gemeinsam sollen die vier Artikel Möglichkeiten aufzeigen, wie Schülerinnen und Schüler dabei unterstützt werden können, ihr kognitives Potential im Physikunterricht umzusetzen.

So geht es im ersten Artikel um die Identifikation von sogenannten Underachievern (Minderleistern), also um Schülerinnen und Schüler, deren Noten in Physik deutlich hinter ihren hohen Intelligenztest-Leistungen zurückliegen. Dieser Artikel beschreibt einen neuen Ansatz zur Identifikation von Underachievern mithilfe einer Analyse latenter Profile. Dieser Ansatz wurde an einer Stichprobe von $N = 316$ Schweizer Gymnasiastinnen und Gymnasiasten erprobt. Die resultierenden Profile erwiesen sich als theoretisch sinnvoll und methodisch valide. Ein Profil von Physik Underachievern, die trotz hohen intellektuellen Potentials unterdurchschnittliche Physiknoten aufwiesen, zeigte sich nur bei den Mädchen. Die als Underachiever klassifizierten Schülerinnen fielen zudem durch geringes Interesse an Physik und ein niedriges fachspezifisches Selbstkonzept auf.

Inwiefern die Bewertungspraxis von Physiklehrpersonen zur Erklärung des hohen Anteils weiblicher Underachiever beitragen kann, wird im zweiten Artikel thematisiert. In einer Online-Studie wurde der Frage nachgegangen, ob Physiklehrpersonen dazu tendieren, die Leistung von Schülerinnen anders zu bewerten als die Leistung von Schülern. Eine Stichprobe von $N = 780$ Physiklehrpersonen aus der Schweiz, aus Österreich und aus Deutschland sollte die immer gleiche Prüfungsantwort einer fiktiven Gymnasiastin oder eines fiktiven Gymnasiasten benoten. Die Antwort enthielt richtige Ansätze, war aber nicht ganz korrekt und liess dadurch Raum für Interpretation. Da von den Lehrpersonen neben dem

Herkunftsland auch das Geschlecht und die Länge der Unterrichtserfahrung erfasst wurden, waren differenzierte Analysen möglich. Tatsächlich zeigte sich ein Gender-Bias zuungunsten von Schülerinnen bei allen schweizerischen und österreichischen sowie den weiblichen deutschen Physiklehrpersonen in der ersten Dekade ihrer Karriere: Die halbrichtige Antwort wurde schlechter bewertet, wenn sie vermeintlich von einer Schülerin stammte. Der Bias verschwand allerdings mit zunehmender Lehrerfahrung. Männliche deutsche Physiklehrpersonen zeigten keinerlei Gender-Bias Effekt. Die Ergebnisse der hier beschriebenen Studie sprechen dafür, dass Schüler in Bewertungssituationen, die Raum für Interpretation lassen, im Zweifelsfall mehr Vorschusslorbeeren erhalten als Schülerinnen. Generell zeigen die Abweichungen zwischen den Lehrpersonen in der Beurteilung der Antwort, dass eine faire Bewertung der Leistungen von Schülerinnen und Schülern nicht ganz einfach ist und der Bewertungsprozess besonderer Aufmerksamkeit bedarf.

Der dritte Artikel befasst sich daher eingehender mit dem Bewertungsprozess und bezieht sich dabei auf die Erfassung des Konzeptverständnisses, während im konventionellen Physikunterricht vor allem quantitatives Problemlösen erhoben wird. Der Artikel beschreibt die Entwicklung und Evaluation des „basic Mechanics Concept Tests“ (bMCT). Der bMCT ist ein an das Physik-Curriculum des Gymnasiums angepasster Multiple-Choice-Test, der grundlegendes konzeptuelles Verständnis in Newtonscher Mechanik erfasst. Die Entwicklung des Tests erfolgte mit einer Stichprobe von $N = 239$ Gymnasiastinnen und Gymnasiasten. Eine anschließende Evaluationsstudie mit $N = 141$ Gymnasiastinnen und Gymnasiasten bestätigte, dass der Test einer Rasch-Skalierung folgt und zur validen Bestimmung von Fortschritten im Konzeptverständnis verwendet werden kann.

Die in den ersten drei Artikeln behandelten Aspekte werden in einer umfassenden Studie zusammengeführt, die im vierten Artikel beschrieben wird. Hinzu kam eine Variation in der Art der Instruktion. In einer Interventionsstudie mit $N = 172$ Gymnasiastinnen und Gymnasiasten wurde untersucht, ob Instruktion, in die gezielt kognitiv aktivierende Elemente einbezogen wurden, die das Konzeptverständnis fördern, für alle Lernende von Vorteil ist und Underachievement entgegenwirken kann. So erhielten vier Schulklassen 18 Lektionen kognitiv aktivierender konzeptbasierter Instruktion und vier weitere Schulklassen 18 Lektionen konventioneller Instruktion in Newtonscher Mechanik. Vier Physiklehrpersonen unterrichteten jeweils eine Klasse auf kognitiv aktivierende Art und Weise und eine Klasse wie gewohnt (d.h., konventionelle Instruktion). Über drei Messzeitpunkte hinweg (Prä, Post und Follow-up) wurden quantitatives Problemlösen und konzeptuelles Verständnis (mithilfe

des bMCTs) zusammen mit motivationalen Variablen erhoben, um die konventionelle mit der kognitiv aktivierenden konzeptbasierten Instruktion zu vergleichen. Die Schülerinnen und Schüler profitierten auf beiden Leistungsmassen von kognitiv aktivierender konzeptbasierter Instruktion, nicht jedoch hinsichtlich der motivationalen Variablen. In latenten Profil-Analysen zeigte sich, dass das Konzeptverständnis weiblicher Underachiever durch kognitiv aktivierende konzeptbasierte Instruktion im Follow-up-Test deutlich angehoben werden konnte, was sich allerdings nicht unmittelbar im Post-Test widerspiegelte und auf die Noten auswirkte.

Aus jeder der vier Studien können eigenständige Implikationen für den Physikunterricht und die Bildungsforschung abgeleitet werden. Gemeinsam ermöglichen sie wichtige Einblicke in das Zusammenspiel von Geschlecht, Underachievement und konzeptbasierter Instruktion im gymnasialen Physikunterricht, die über den Rahmen jedes einzelnen Artikels hinausgehen. In einer Zusammenschau der Befunde werden abschliessend drei übergeordnete Themenbereiche diskutiert: Faktoren, die zu Underachievement beitragen könnten, Wege zum Abbau der Geschlechtsunterschiede in Physik sowie die generelle Gestaltung des Physikunterrichts. Die dabei abgeleiteten allgemeinen Empfehlungen für zukünftige Forschung und Bildungspraxis zeigen Möglichkeiten auf, einer grösseren Anzahl von Lernenden, Schülerinnen und Schülern, zu einem besseren Verständnis physikalischer Zusammenhänge und einer erfolgreichen Investition ihres kognitiven Potentials zu verhelfen.

Table of Contents

Abstract.....	I
Zusammenfassung.....	IV
1. General Introduction	1
Instruction in the Physics Classroom	3
The Student in the Physics Classroom.....	4
Assessment in the Physics Classroom.....	4
Introducing the Four Papers of this Thesis	6
What this Thesis Aims to Achieve.....	12
References.....	15
2. Underachievement in Physics: When Intelligent Girls Fail	21
Introduction.....	22
Operational Definitions of Underachievement	23
Motivational Correlates of Physics Underachievement	25
Gender Differences and Physics Underachievement	26
The Present Study	27
Method	27
Results.....	35
Discussion	44
References.....	48
Acknowledgements.....	55
3. Studying Gender Bias in Physics Grading: The Role of Teaching Experience and Country ..	56
Introduction.....	57
Gender Bias in Teachers' Judgments in STEM Fields	57
The Present Study	61
Method	62
Results.....	67
Discussion	75
Conclusion	78
References.....	80
Acknowledgements.....	84
4. The basic Mechanics Concept Test (bMCT): An Efficient Rasch-Scaled Multiple Choice Test of Fundamental Conceptual Understanding in Newton's Mechanics	85
Introduction.....	86
Conceptual Knowledge	87

The Seminal Role of the FCI	87
Shortcomings of Existing Concept Tests in Newton's Mechanics	88
A New Instrument.....	89
Method	90
Results.....	99
Discussion	104
References.....	108
Acknowledgements.....	112
5. Fostering Conceptual Understanding with Cognitively Activating Instruction in Physics Classrooms: Evidence for General Effects and Special Benefits for High Potential Students ..	113
Introduction.....	114
Conceptual Knowledge and Conceptual Learning in Physics	116
Effectiveness of Conceptual Instruction: Conceptual and Procedural Knowledge.....	117
Cognitively Activating Instructional Methods.....	120
Motivation in Cognitively Activating Physics Learning.....	123
The Problem of Physics Underachievement	124
The Present Study	125
Method	128
Results.....	144
Discussion	155
Conclusion	170
References.....	171
Acknowledgements.....	183
6. General Discussion.....	184
Integrative Summary of the Main Findings	185
What Factors May Contribute to the Underachievement of Some Female Students in Secondary School Physics Classrooms?	187
What is the Contribution of the Present Work Regarding the Gender Gap in Physics?	191
What is the Contribution of the Present Work Regarding the Design of Physics Lessons?	195
References.....	199
Appendix A: Motivational Scales	202
Appendix B: bMCT (plus).....	206
Appendix C: Quantitative Problem Solving Test.....	218

1. General Introduction

Understanding physics enables adolescents not only to choose from a broad variety of professional careers. Understanding physics enables them to explain what is happening in the physical world around them. Yet, even after having attended many lessons of physics instruction, a substantial part of students lacks knowledge of basic physics concepts, as repeatedly shown in the field of Newtonian mechanics (e.g., Beaton et al., 1996; Halloun & Hestenes, 1985; Hestenes, Wells, & Swackhamer, 1992; McDermott, 1984; Nieminen, Savinainen, & Viiri, 2010). The students' naïve explanations of how things work often interfere with the scientifically accepted explanations teachers are trying to transmit (see e.g., Carey, 2000; Ohlsson, 2013; Smith III, diSessa, & Roschelle, 1994; Vosniadou, Ioannides, Dimitrakopoulou, & Papademetriou, 2001). *So what can be done to help secondary school students to invest their intellectual potential in understanding physics?* Attempting to provide a solution, it is crucial to determine which factors (e.g., teachers, instruction, genetics) have to be considered. This also involves deciding on the level of granularity required to find a satisfactory answer to this question.

There is extensive research addressing diverse aspects related to learning in the domain of physics or STEM (science, technology, engineering, and mathematics) disciplines more generally (see e.g., Newcombe et al., 2009; Redish, 2004). Researchers have focused on cognitive abilities predicting physics learning (e.g., Wai, Lubinski, & Benbow, 2009), on the cognitive processes involved in developing physics knowledge (e.g., DiSessa, 1993), on designing effective instructional tools (e.g., Clark & Jorde, 2004), on evaluating classroom interventions (e.g., Crouch & Mazur, 2001), on implementing teacher professional development programs (e.g., Ostermeier, Prenzel, & Duit, 2010), on analyzing classroom discourse and interaction (e.g., van Zee & Minstrell, 1997), on the influence of motivational variables, beliefs, and expectations (e.g., Jansen, Schroeders, & Lüdtke, 2014), on the role of assessment (e.g., Dufresne & Gerace, 2004), or on supporting and debilitating structures outside school (e.g., Harackiewicz, Rozek, Hulleman, & Hyde, 2012). The present thesis focuses on physics learning as it happens – or does not happen – in the physics classroom. This level of analysis allows to directly research the environment where learning primarily takes place. Focusing on the most prominent factors operating in the classroom (and not beyond) enables the researcher to immediately intervene and potentially access all relevant factors in the sense of an engineering education approach as advocated by design researchers

(see Brown, 1992; Collins, Joseph, & Bielaczyc, 2004). In four papers, this thesis compiles different lines of research related to physics learning that are considered particularly relevant for investigating physics learning on the level of the physics classroom. It concentrates on the three factors *instruction*, *the student*, and *assessment*, as well as their *interactions* to comprehensively investigate the immediate learning conditions in secondary school physics classrooms aiming at the deduction of sound practical measures.

To achieve this, two main samples are examined: a student sample and a teacher sample. In the student sample, data from $N = 418$ Swiss secondary school students from the highest track of the Swiss educational system, the Gymnasium, were collected. In addition to personal information (including gender and age), physics grades, measures of intelligence and conceptual understanding (using a test described in the third paper), and motivational variables were gathered from all of the students. In parts, students were recruited individually by advertising and contacting teachers and student representations. This part of the sample comprises $n = 133$ students. The second part, $n = 285$ students, consists of 14 whole physics classrooms that participated in an intervention study. Supported by the MINT-Learning Center of the ETH Zurich, six of the 14 classes were examined in the context of a pilot study and eight classes took part in the main intervention study that is described in the fourth paper. The first paper that presents a correlational study and the third paper on the development and evaluation of a test instrument are both based on parts of the whole student sample.

The teacher sample consists of $N = 780$ German-speaking secondary school physics teachers from Switzerland, Austria, and Germany. Teacher data, including personal information, like years of teaching experience and gender, and a grade that the teachers assigned to a fictive student test answer, were collected using an online-survey tool. The teacher sample is analyzed in the context of an experimental study that is described in the second paper. The four papers that constitute this thesis are introduced in more detail later in this general introduction.

As substantiated in the following sections, a rather coarse-grained level of analysis is chosen to investigate characteristics of instruction, while particular characteristics of the students and the process and focus of assessment are examined. Subsequently, the S(tudent)I(nstruction)A(ssessment)-Interaction-Framework is used to introduce and embed the four papers that constitute the thesis. Integrating the four papers within the framework, *the interplay between gender, underachievement, and conceptual instruction* emerges as

central issue. The introduction closes with the specific research aims that are addressed in each of the four papers and the overarching research aims that are addressed by considering the overall picture resulting from a synopsis of the four parts of this thesis.

Instruction in the Physics Classroom

In the present work, physics instruction is analyzed on the macro-level, in the sense that this thesis does not investigate the isolated effectiveness of single instructional methods or principles but the general instructional orientation reflected in prolonged physics instruction. Many instructional tools and methods that educational researchers developed during the last decades are based on the same conception of learning that emphasizes the learner's active construction of new conceptual knowledge based on already existing knowledge (see Berthold & Renkl, 2010; Schneider & Stern, 2010a). Conceptual knowledge can be described as abstract and general knowledge of a domain's main principles and their connections (Carey, 2000; Schneider & Stern, 2010b). In the domain of physics, this kind of knowledge may exemplarily include understanding of the concepts "body movement in response to forces" or "momentum conservation" (see Halloun & Hestenes, 1985). Instruction translating this "social-constructivist" perspective on learning can be considered promising in promoting learning in physics and in STEM disciplines in general (see McDermott, 1984; Rosenquist & McDermott, 1987). The success of a number of cognitively activating methods that all utilize this general idea, such as prompting students to generate self-explanations (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Schworm & Renkl, 2007) or confronting students' existing mental models with experts' models (Gadgil, Nokes-Malach, & Chi, 2012), as well as the success of whole instructional approaches that are oriented towards this general idea (e.g., Baumert et al., 2010; Freeman et al., 2014; Lorenzo, Crouch, & Mazur, 2006; Staub & Stern, 2002) point to the high relevance of what may also be called "conceptual instruction" for advancing STEM education. In contrast to conceptual instruction, conventional physics instruction can be expected to leave less room for actively working on the conceptual knowledge that underlies formulae and take home messages (Langer Tesfaye & White, 2012; Seidel et al., 2006; Taconis, 1995; Zohar & Sela, 2003). To sum up, this thesis contrasts conventional instruction with conceptual instruction that is highly promising in fostering student learning in physics.

The Student in the Physics Classroom

Instead of only addressing students on average, learners with specific characteristics may require particular attention in the context of physics learning. Consequently, to investigate physics learning, specific student characteristics are considered in interaction with characteristics of the instruction. There is broad evidence that especially female students struggle with physics (e.g., Beaton et al., 1996; Lubinski & Benbow, 1992; Organisation for Economic Co-operation and Development, 2009; Taasoobshirazi & Carr, 2008). Boys also seem to be equipped with a more beneficial motivational background, demonstrated, for example, with regard to physics self-concept (e.g., Debacker & Nelson, 2000; Jansen et al., 2014) or interest in physics (e.g., Adams et al., 2006; Hoffmann, 2002). In addition to gender, the students' intelligence has to be taken into account. Effective instruction should stimulate the students to deploy their intellectual potential in the physics classroom. Consequently, learners whose physics achievement considerably lags behind their intellectual potential have to be examined carefully. Combined with the findings on the gender gap in physics, there is reason to expect more intelligent females than males to underachieve in the domain of physics (c.f. Lubinski & Benbow, 1992). Conceptual instruction, however, has the potential to particularly activate this group of students, fitting their needs (c.f. Häussler & Hoffmann, 2002; Hulleman & Harackiewicz, 2009; Zohar & Sela, 2003). To sum up, while this thesis focuses specifically on investigating how the student characteristics gender and intelligence influence learning from conventional or conceptual physics instruction, motivational variables are considered in addition to obtain a more detailed picture of particular groups of learners.

Assessment in the Physics Classroom

Besides characteristics of the instruction and the student, the assessment has to be taken into account. Although assessment is able to fulfill diverse functions (when used in a formative way, for instance), at school, assessment usually serves the purpose of assigning grades (in the sense of a summative assessment). The outcomes of assessment at school (i.e., grades) considerably determine the students' future opportunities, provide the most tangible feedback on their capabilities, and are closely intertwined with the students' academic interests, self-concepts, future school-engagement, and school achievement (e.g.,

Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Poorthuis et al., 2014; von Maurice, Dörfler, & Artelt, 2014). The feedback itself hence affects future performance, often mediated by motivational variables, and may trigger self-fulfilling prophecy (e.g., de Boer, Bosker, & van der Werf, 2010; Jussim & Eccles, 1992) or stereotype threat effects (e.g., Marchand & Taasobshirazi, 2013; Nguyen & Ryan, 2008). Moreover, the focus of assessment can be expected to directly influence what students learn from the instruction, when they study what the assessment requires them to learn in order to pass the examination (see Boud & Falchikov, 2007; Gibbs & Simpson, 2004). The kinds of competencies that are commonly required to succeed in physics assessments may be interpreted as the kinds of competencies that are generally regarded as important to accomplish in the course of physics instruction (see National Research Council, 2001). The fit between certain student characteristics and the competencies emphasized in the assessment may further influence learning. There is reason to assume, for instance, that manipulating and applying formulae is more appealing to male students than to female students (see e.g., Kang & Wallace, 2005; Taconis, Ferguson-Hessler, & Broekkamp, 2001; Zohar, 2006; Zohar & Sela, 2003). Considering the possibility of gender biased grading, the process of the assessment may also affect the students, depending on their gender. Gender-STEM stereotypes may influence how teachers evaluate the performance of female vs. male students (see Heller, Finsterwald, & Ziegler, 2010; Miller, Eagly, & Linn, 2014; Nosek et al., 2007; Nosek, Banaji, & Greenwald, 2002). To sum up, the process and focus of assessment can be construed as an interacting transmitter between characteristics of the instruction and characteristics of the student. Assessment, moreover, is meant to reflect (to differing degrees, depending on the process and focus of the assessment) the result of the learning process that, in turn, is regarded as a function of the interplay between instruction and the student.

According to these considerations, Figure 1.1 shows the SIA-Interaction-Framework of Physics Learning (SIA is an acronym for student, instruction, and assessment) that results from connecting the three factors based on their theoretical interrelations. Learning, as the desired outcome, may result from individual students receiving a certain kind of instruction (depicted by the straight double arrow connecting instruction and student). Learning is assessed, more or less accurately (depicted by the straight arrow from ‘the learning arrow’ to assessment). In the whole process, however, learning is also indirectly influenced by the process and focus of the assessment itself. The outcome of assessment that can impact on

student characteristics such as learning amotivation or the domain-specific self-concept, can itself be expected to partially depend on the interaction between individual student characteristics and the process and focus of the assessment (e.g., gender bias in grading, individual students' aversion to only re-arranging equations). The focus of assessment, in turn, can also affect what a student focuses on during instruction. This is why assessment is represented by a thick double arrow in Figure 1.1 that interacts with and transmits between characteristics of the student and characteristics of the instruction.

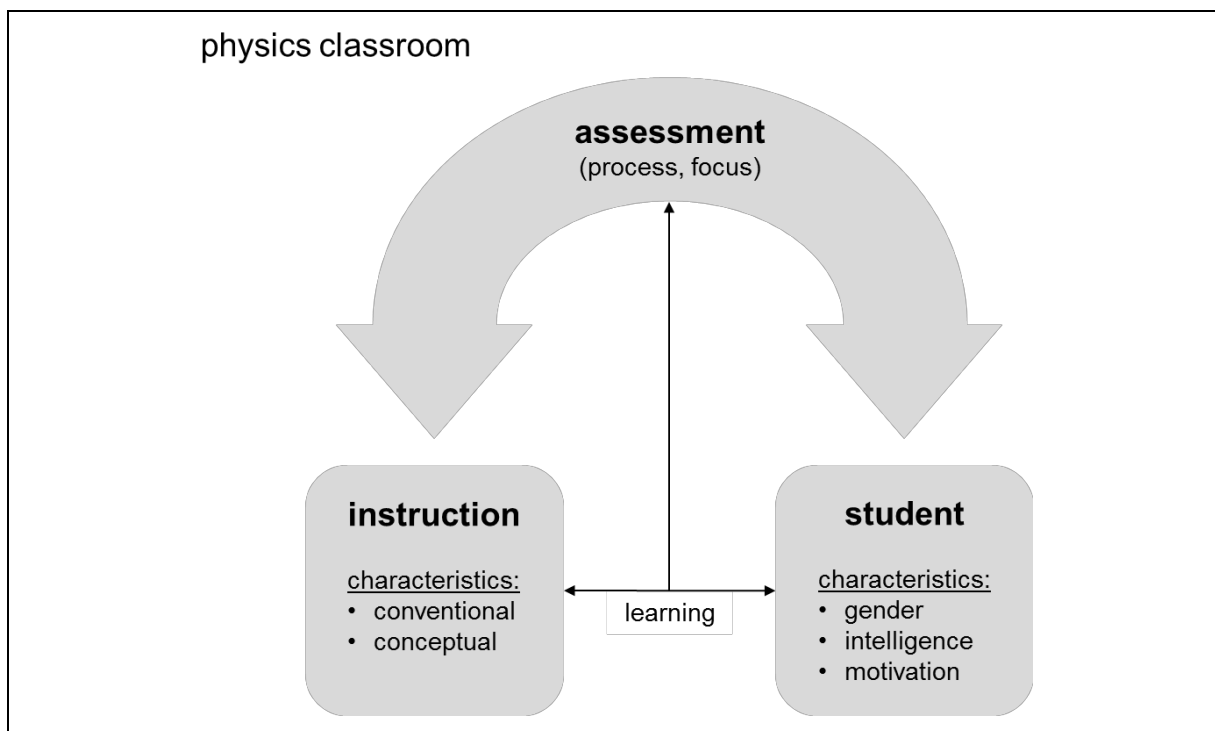


Figure 1.1. The SIA-Interaction-Framework of Physics Learning that visualizes learning as a function of the student (S), instruction (I), and assessment (A), as well as their interactions in the physics classroom.

Introducing the Four Papers of this Thesis

The SIA-Interaction-Framework is used to illustrate the scope of this thesis. Hence, the four papers that constitute this thesis are introduced in the following sections referring to their localization within the framework (see Figure 1.2).

Underachievement in Physics: When Intelligent Girls Fail

The first paper focuses on the interaction between the student characteristics gender and intelligence (and, in addition, motivational variables), on the one hand, and assessment, on the other hand (see upper left part of Figure 1.2). In this paper, assessment is investigated on a very general level without going into detail in terms of the process or focus of assessment. The most common outcomes of assessment in the classroom, i.e., grades, are used to take a look at the relationship between assessment and the students' gender and intellectual potential. This first paper provides a cross-sectional status check describing how the students' intelligence is reflected in their physics grades as a function of the students' gender. It examines the assumption that more intelligent female than male students underachieve in the domain of physics, i.e., receive bad physics grades despite a high intellectual potential. In doing so, this first study presents a sound rationale for considering the student characteristics gender and intelligence when investigating physics learning within the complex dynamics of the physics classroom. The first paper with the title "Underachievement in Physics: When Intelligent Girls Fail" that is submitted for publication is summarized in the following paragraph.

The present study examined gender-specific physics underachievement to identify highly intelligent female and male students who perform below their intellectual potential in physics. The sample consisted of 316 students (182 girls) from higher secondary school (Gymnasium) in Switzerland (age $M = 16.25$ years, $SD = 1.12$ years). In a multiple group latent profile analysis, intellectual potential and physics grades were used to determine gender-specific student profiles. In accordance with prior expectations, a problematic profile of female physics underachievers with high intellectual potential but below average physics grades was identified. Their math grades and GPA, by contrast, turned out to be within the normal range suggesting domain-specific underachievement. The female physics underachievers, moreover, showed a low interest and self-concept in physics compared to the other students, complementing the picture. An independent sample was used to validate the student profiles. We finally discuss implications for physics classrooms and future research.

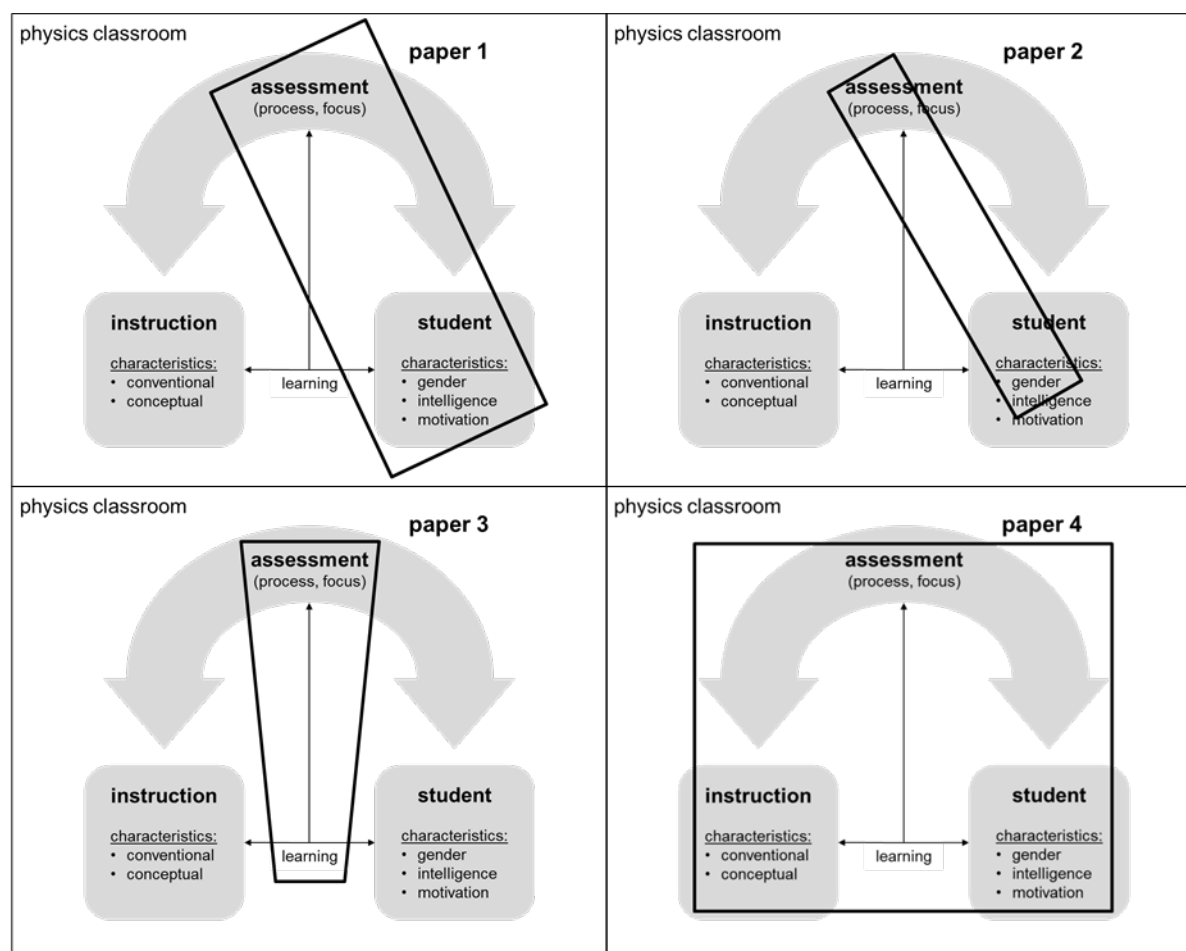


Figure 1.2. Localization of the four papers of this thesis within the SIA-Interaction-Framework of Physics Learning.

Studying Gender Bias in Physics Grading: The Role of Teaching Experience and Country

Inspired by the results of the first paper and, in particular, the detection of pronounced physics underachievers among female secondary school students, the second study elaborates on the interaction between the student and the assessment by zooming in on the assessment process. The considerably low performance of some intelligent girls in terms of physics grades may partially result from a general gender bias in physics teachers' grading. Consequently, the second paper investigates the process of assessment as a function of the students' gender (see upper right part of Figure 1.2). The abstract of the second paper entitled

“Studying Gender Bias in Physics Grading: The Role of Teaching Experience and Country” that is submitted for publication is provided below.

The existence of gender-STEM (science, technology, engineering, and mathematics) stereotypes has been repeatedly documented. This article examines physics teachers’ gender bias in grading and the influence of teaching experience in Switzerland, Austria, and Germany. In a 2×2 between-subjects design, with years of teaching experience included as moderating variable, physics teachers ($N = 780$) from Switzerland, Austria, and Germany graded a fictive student’s answer to a physics test question. While the answer was exactly the same for each teacher, only the student’s gender and specialization in languages vs. science were manipulated. Specialization was included to gauge the relative strength of potential gender bias effects. Multiple group regression analyses, with the grade that was awarded as the dependent variable, revealed only partial cross-border generalizability of the effect pattern. While the overall results in fact indicated the existence of a consistent and clear gender bias against girls in the first part of physics teachers’ careers that disappears with increasing teaching experience for Swiss teachers, Austrian teachers, and German female teachers, German male teachers showed no gender bias effects at all. The results are discussed regarding their relevance for educational practice and research.

The basic Mechanics Concept Test (bMCT): An Efficient Rasch-Scaled Multiple Choice Test of Fundamental Conceptual Understanding in Newton’s Mechanics

The findings of the first two papers warrant a closer examination of both the process and focus of assessment that is meant to reflect the results of a learning process (see lower left part of Figure 1.2). Despite the only partial generalizability of gender bias effects, gender bias in grading seems to represent a real problem in at least some physics classes (second paper). Biased assessment may contribute to the underachievement of some girls (as uncovered in the first paper), both directly by distorting the achievement measure and indirectly by negatively affecting motivation related to physics. Consequently, the third paper provides a measure that can validly gauge learning and the resulting level of understanding in Newton’s mechanics. Such a measure is necessary to derive sound conclusions about the effectiveness of instruction and is hence a prerequisite for the intervention study described in the fourth paper.

Regarding the process of assessment, the measurement has to ascertain high objectivity to avoid biased evaluations of student answers. This can be achieved by designing a test that conforms to a strict measurement model (e.g., the Rasch model, see Rasch, 1960). The focus of assessment requires consideration, too. Conventional physics examinations usually do not aim at eliciting flexible conceptual knowledge that enables students to participate in scientific discourses and tackle new problem situations but rather ask students to remember facts and solve convergent quantitative problems (c.f. Alberts, 2009; Langer Tesfaye & White, 2012). Female learners may suffer particularly from assessments that focus on formulae application and the recall of detached abstract facts (see Boaler, 1997; Zohar, 2006). Assessment focusing on conceptual knowledge may hence provide an informative alternative, measuring a highly relevant learning outcome, conceptual understanding. Conceptual knowledge is a prerequisite for flexible, context-independent problem solving (see Hiebert, 1986) which is considered a key element of physics literacy (McDermott, 1984; Resnick, 2010). In the end, the instrument also has to be easily applicable in the physics classroom and allow efficient assessment of relevant knowledge. The development and evaluation of a gender-fair assessment instrument conforming to all of these requirements is described in the third paper entitled “The basic Mechanics Concept Test (bMCT): An Efficient Rasch-Scaled Multiple Choice Test of Fundamental Conceptual Understanding in Newton’s Mechanics” that is under second review. The paper is summarized in the following.

Solid assessment of understanding in Newton’s mechanics is highly relevant both for physics classrooms and research. Several concept tests have been developed. What is still missing, however, is an efficient test that is adapted to the content taught to secondary school students and that can be validly applied as pre- and posttest to reflect learning progress. In this paper, we describe the development and evaluation of the basic Mechanics Concept Test (bMCT) that was designed to meet these requirements. In the context of test development, qualitative as well as quantitative methods including Rasch analyses were applied to a sample of $N = 239$ Swiss secondary school students. The final test’s conformity to the Rasch model was confirmed with a sample of $N = 141$ students. We further ascertained the bMCT’s applicability as change measure. Additionally, the criterion validity of the bMCT and the Force Concept Inventory (FCI) was compared in a sample of secondary school students ($N = 66$) and a sample of mechanical engineering students ($N = 21$). In both samples, the bMCT clearly outperformed the FCI in predicting actual student

performance. The paper closes with a discussion on the bMCT's potential regarding physics education and research purposes.

Fostering Conceptual Understanding with Cognitively Activating Instruction in Physics Classrooms: Evidence for General Effects and Special Benefits for High Potential Students

The three papers introduced so far are all concerned with the student, assessment, and their interactions. The factor instruction has not been explicitly addressed. In the fourth paper, conceptual physics instruction is implemented and contrasted with conventional instruction that very generally represents a physics teacher's normal instruction. The information gained and work done in the first three papers guide the analytic strategy of this final study. The last part of this thesis hence combines the two factors student and assessment with characteristics of the instruction within one comprehensive final study (see lower right part of Figure 1.2). The fourth paper investigates the potential of cognitively activating conceptual instruction to address deficient physics knowledge in general and female students' underachievement in particular, considering alternative assessments of performance (the bMCT) in addition to physics grades. The next paragraph summarizes the fourth paper with the title "Fostering Conceptual Understanding with Cognitively Activating Instruction in Physics Classrooms: Evidence for General Effects and Special Benefits for High Potential Students" that is to be submitted for publication.

Secondary school physics instruction is confronted with the students' deficient physics literacy in general and some girls' underachievement in particular. In this study, we investigate the potential of cognitively activating (CogAct) physics instruction that is focused on conceptual understanding to address these two issues. While positive effects of single CogAct instructional elements have already been confirmed, little is known about the effectiveness of a whole CogAct teaching unit implemented in physics classrooms. Four teachers participated with two classes each. They taught one of their classes based on an 18-lessons unit of CogAct instruction, while the other class was instructed as always with instructional time and content matched. Across three points of measurement, we gathered measures of conceptual understanding, quantitative problem solving, and motivation of $N = 172$ (92 girls) Swiss secondary school students. The results of multiple regression analyses showed

that CogAct instruction was superior to conventional instruction in terms of both performance measures. CogAct students, however, required a conceptual scaffold at the quantitative problem solving posttest in order to outperform conventional students at follow-up problem solving. The advantages of CogAct instruction were not reflected in any of the motivational variables in the overall sample. Additional latent profile analyses revealed that underachieving girls, high achieving boys, and, particularly, high achieving girls profited considerably from CogAct instruction regarding performance and motivation. We discuss the findings of the present study in the light of the potential of transformed instruction and assessment to promote effective and gender-fair physics learning.

What this Thesis Aims to Achieve

In the remainder of the introduction, the specific aims that guide the research described in each of the four papers are summarized. In addition, three overarching research questions that are addressed by considering the overall picture are formulated.

The Four Papers

First paper. The first paper “Underachievement in Physics: When Intelligent Girls Fail” aims at investigating underachievement in physics in terms of its gender-specific prevalence expecting especially girls to underachieve.

Second paper. The second paper “Studying Gender Bias in Physics Grading: The Role of Teaching Experience and Country” examines gender bias in physics teachers’ grading. The study further investigates the potential moderating effect of teaching experience, which may reduce gender bias with increasing years of practice.

Third paper. The aim of the work described in the third paper “The basic Mechanics Concept Test (bMCT): An Efficient Rasch-Scaled Multiple Choice Test of Fundamental Conceptual Understanding in Newton’s Mechanics” is to develop and evaluate a test that objectively assesses conceptual knowledge in basic Newtonian mechanics, that is efficient,

that is adapted to the content taught to secondary school students, and that can be validly applied both as pre- and posttest to reflect learning progress.

Fourth paper. In the fourth paper “Fostering Conceptual Understanding with Cognitively Activating Instruction in Physics Classrooms: Evidence for General Effects and Special Benefits for High Potential Students”, the following four research questions are addressed:

1) *General effectiveness.* Is conceptual instruction beneficial for all students in terms of a conceptual transfer measure when compared to conventional instruction in physics classrooms? Does conceptual instruction also prove beneficial for the acquisition of procedural problem solving skills?

2) *Accessing procedures via concepts.* Does a conceptual scaffold help students who had received conceptual instruction to access quantitative problem solving procedures via their conceptual knowledge? Is it less beneficial for conventional learners who may have developed a less elaborated conceptual knowledge base that less strongly connects to quantitative problem solving procedures?

3) *Impact on motivation.* Does conceptual instruction increase the students’ interest in physics, physics self-concept, and use of efficient learning strategies, as well as decrease the students’ learning amotivation and physics anxiety as compared to conventional physics instruction?

4) *Impact on physics underachievement.* Given that physics underachievement is defined by the systematic co-occurrence of high intellectual potential and low physics grades, can conceptual physics instruction tackle underachievement?

Synopsis

In a synopsis of the findings of this thesis, the interaction between the student, assessment, and instruction can be specified in terms of the students’ gender, intelligence, and motivation, the process and focus of the assessment, and the instruction’s focus. In conjunction, the four papers allow drawing general conclusions concerning the immediate learning conditions in secondary school physics classrooms that go beyond the scope of each

individual paper. Taken together, they provide important insights into the interplay between gender, underachievement, and conceptual instruction in physics. So in addition to the more specific research questions dealt with in each of the four papers, this thesis aims at providing sound answers to three overarching research questions addressing more general issues:

1. What factors may contribute to the underachievement of some female students in secondary school physics classrooms?
2. What is the contribution of the present work regarding the gender gap in physics?
3. What is the contribution of the present work regarding the design of physics lessons?

In a nutshell, this thesis is intended to figure out what can be done to enable secondary school students, both female and male, to invest their intellectual potential in understanding physics by adding new perspectives and findings that emphasize the interplay between gender, underachievement, and conceptual instruction. The main body of this thesis consists of the four papers that include two (quasi-)experimental studies (the second and fourth paper) and one correlational study (the first paper), as well as the development and evaluation of a concept test (the third paper). The papers are presented in the order described, followed by the final chapter of this thesis, the general discussion. After a summary of the main findings in light of the SIA-Interaction-Framework, the general discussion turns to the three overarching research questions. The thesis' contribution to answer these questions is reviewed and recommendations for educational practice and potential starting points for future research are derived.

In line with Bandura who emphasized that “[b]ehavior, cognitive and other personal factors, and environmental influences all operate interactively as determinants of each other” (1986, p. 23) and a social cognitive approach to the study of student learning, the work described in the following investigates how students achieve and underachieve, are evaluated, and, ultimately, understand and learn as a function of their gender and intelligence, the assessment, and the instruction in the physics classroom.

References

- Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 2(1), 1–14. <http://doi.org/10.1103/PhysRevSTPER.2.010101>
- Alberts, B. (2009). Redefining science education. *Science*, 323(5913), 437–437. <http://doi.org/10.1126/science.1170933>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180. <http://doi.org/10.3102/0002831209345157>
- Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Berthold, K., & Renkl, A. (2010). How to foster active processing of explanations in instructional communication. *Educational Psychology Review*, 22(1), 25–40. <http://doi.org/10.1007/s10648-010-9124-9>
- Boaler, J. (1997). Reclaiming school mathematics: The girls fight back. *Gender and Education*, 9(3), 285–305. <http://doi.org/10.1080/09540259721268>
- Boud, D., & Falchikov, N. (Eds.). (2007). *Rethinking assessment in higher education: Learning for the longer term*. Routledge.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2(2), 141–178.
- Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, 21(1), 13–19. [http://doi.org/10.1016/S0193-3973\(99\)00046-5](http://doi.org/10.1016/S0193-3973(99)00046-5)
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182. http://doi.org/10.1207/s15516709cog1302_1
- Clark, D., & Jorde, D. (2004). Helping students revise disruptive experientially supported ideas about thermodynamics: Computer visualizations and tactile models. *Journal of Research in Science Teaching*, 41(1), 1–23. <http://doi.org/10.1002/tea.10097>

- Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, 13(1), 15–42.
- Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970–977.
- Debacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *The Journal of Educational Research*, 93(4), 245–254. <http://doi.org/10.1080/00220670009598713>
- De Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1), 168–179. <http://doi.org/10.1037/a0017289>
- DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2-3), 105–225.
- Dufresne, R. J., & Gerace, W. J. (2004). Assessing-to-learn: Formative assessment in physics instruction. *The Physics Teacher*, 42(7), 428–433.
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. <http://doi.org/10.1073/pnas.1319030111>
- Gadgil, S., Nokes-Malach, T. J., & Chi, M. T. H. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction*, 22(1), 47–61. <http://doi.org/10.1016/j.learninstruc.2011.06.002>
- Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, 1(1), 3–31.
- Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53(11), 1056–1065. http://doi.org/10.1007/978-3-642-20072-4_12
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, 100(1), 105–122. <http://doi.org/10.1037/0022-0663.100.1.105>
- Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, 23(8), 899–906.
- Häussler, P., & Hoffmann, L. (2002). An intervention study to enhance girls' interest, self-concept, and achievement in physics classes. *Journal of Research in Science Teaching*, 39(9), 870–888. <http://doi.org/10.1002/tea.10048>

- Heller, K. A., Finsterwald, M., & Ziegler, A. (2010). Implicit theories of mathematics and physics teachers on gender-specific giftedness and motivation. In K. A. Heller (Ed.), *Munich studies of giftedness* (pp. 239–252). Berlin: LIT.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141–158. <http://doi.org/10.1119/1.2343497>
- Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.
- Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction*, 12(4), 447–465. [http://doi.org/10.1016/S0959-4752\(01\)00010-X](http://doi.org/10.1016/S0959-4752(01)00010-X)
- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, 326(5958), 1410–1412.
- 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. <http://doi.org/10.1016/j.lindif.2013.12.003>
- Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, 63(6), 947–961.
- Kang, N.-H., & Wallace, C. S. (2005). Secondary science teachers' use of laboratory activities: Linking epistemological beliefs, goals, and practices. *Science Education*, 89(1), 140–165. <http://doi.org/10.1002/sce.20013>
- Langer Tesfaye, C., & White, S. (2012). *High school physics teacher preparation*. American Institute of Physics Statistical Research Center.
- Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2), 118–122. <http://doi.org/10.1119/1.2162549>
- Lubinski, D., & Benbow, C. P. (1992). Gender differences in abilities and preferences among the gifted: Implications for the math-science pipeline. *Current Directions in Psychological Science*, 1(2), 61–66. <http://doi.org/10.1111/1467-8721.ep11509746>
- Marchand, G. C., & Taasobshirazi, G. (2013). Stereotype threat and women's performance in physics. *International Journal of Science Education*, 35(18), 3050–3061. <http://doi.org/10.1080/09500693.2012.683461>
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. <http://doi.org/10.1111/j.1467-8624.2005.00853.x>
- McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, 37(7), 24–32. <http://doi.org/10.1063/1.2916318>
- Miller, D. I., Eagly, A. H., & Linn, M. C. (2014). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000005>

- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. (J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC: National Academy Press.
- Newcombe, N. S., Ambady, N., Eccles, J., Gomez, L., Klahr, D., Linn, M., ... Mix, K. (2009). Psychology's role in mathematics and science education. *American Psychologist*, 64(6), 538–550. <http://doi.org/10.1037/a0014813>
- Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93(6), 1314–1334.
- Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics - Physics Education Research*, 6(2), 1–12. <http://doi.org/10.1103/PhysRevSTPER.6.020109>
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math \neq me. *Journal of Personality and Social Psychology*, 83(1), 44–59. <http://doi.org/10.1037/0022-3514.83.1.44>
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1), 36–88. <http://doi.org/10.1080/10463280701489053>
- Ohlsson, S. (2013). Beyond evidence-based belief formation: How normative ideas have constrained conceptual change research. *Frontline Learning Research*, 1(2), 70–85. <http://doi.org/10.14786/flr.v1i2.58>
- Organisation for Economic Co-operation and Development (2009). *Top of the class: High performers in science in PISA 2006*. OECD Publishing. Retrieved from <http://www.oecd-ilibrary.org/docserver/download/9809061e.pdf?expires=1394711955&id=id&accname=ocid72024074a&checksum=FBF7E1D665EB5EFDB3C197D19CE9D2D7>
- Ostermeier, C., Prenzel, M., & Duit, R. (2010). Improving science and mathematics instruction: The SINUS project as an example for reform as teacher professional development. *International Journal of Science Education*, 32(3), 303–327. <http://doi.org/10.1080/09500690802535942>
- Poorthuis, A. M. G., Juvonen, J., Thomaes, S., Denissen, Jaap J. A., Orobio de Castro, Bram, & van Aken, Marcel A. G. (2014). Do grades shape students' school engagement? The psychological consequences of report card grades at the beginning of secondary school. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000002>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

- Redish, E. F. (2004). A theoretical framework for physics education research: Modeling student thinking. *arXiv Preprint physics/0411149*. Retrieved from <http://arxiv.org/abs/physics/0411149>
- Resnick, L. B. (2010). Nested learning systems for the thinking curriculum. *Educational Researcher*, 39(3), 183–197.
- Rosenquist, M. L., & McDermott, L. C. (1987). A conceptual approach to teaching kinematics. *American Journal of Physics*, 55(5), 407–415. <http://doi.org/10.1119/1.15122>
- Schneider, M., & Stern, E. (2010a). The cognitive perspective on learning: Ten cornerstone findings. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 69–90). Paris: OECD Publishing.
- Schneider, M., & Stern, E. (2010b). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology*, 46(1), 178–192. <http://doi.org/10.1037/a0016701>
- Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology*, 99(2), 285–296.
- Seidel, T., Prenzel, M., Rimmele, R., Dalehefte, I. M., Herweg, C., Kobarg, M., & Schwindt, K. (2006). Blicke auf den Physikunterricht. Ergebnisse der IPN Videostudie. *Zeitschrift Für Pädagogik*, 52(6), 799–821.
- Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2), 115–163. http://doi.org/10.1207/s15327809jls0302_1
- Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, 94(2), 344–355. <http://doi.org/10.1037/0022-0663.94.2.344>
- Taasoobshirazi, G., & Carr, M. (2008). Gender differences in science: An expertise perspective. *Educational Psychology Review*, 20(2), 149–169. <http://doi.org/10.1007/s10648-007-9067-y>
- Taconis, R. (1995). *Understanding based problem solving: Towards qualification-oriented teaching and learning in physics education*. Technische Universiteit Eindhoven.
- Taconis, R., Ferguson-Hessler, M. G. M., & Broekkamp, H. (2001). Teaching science problem solving: An overview of experimental work. *Journal of Research in Science Teaching*, 38(4), 442–468. <http://doi.org/10.1002/tea.1013>
- Van Zee, E. H., & Minstrell, J. (1997). Reflective discourse: Developing shared understandings in a physics classroom. *International Journal of Science Education*, 19(2), 209–228.
- Von Maurice, J., Dörfler, T., & Artelt, C. (2014). The relation between interests and grades: Path analyses in primary school age. *International Journal of Educational Research*, 64, 1–11. <http://doi.org/10.1016/j.ijer.2013.09.011>

- Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction*, 11(4), 381–419.
- Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4), 817–835. <http://doi.org/10.1037/a0016127>
- Zohar, A. (2006). Connected knowledge in science and mathematics education. *International Journal of Science Education*, 28(13), 1579–1599. <http://doi.org/10.1080/09500690500439199>
- Zohar, A., & Sela, D. (2003). Her physics, his physics: Gender issues in Israeli advanced placement physics classes. *International Journal of Science Education*, 25(2), 245–268. <http://doi.org/10.1080/09500690210126766>

2. Underachievement in Physics: When Intelligent Girls Fail

Sarah I. Hofer and Elsbeth Stern

The present study examined gender-specific physics underachievement to identify highly intelligent female and male students who perform below their intellectual potential in physics. The sample consisted of 316 students (182 girls) from higher secondary school (Gymnasium) in Switzerland (age $M = 16.25$ years, $SD = 1.12$ years). In a multiple group latent profile analysis, intellectual potential and physics grades were used to determine gender-specific student profiles. In accordance with prior expectations, a problematic profile of female physics underachievers with high intellectual potential but below average physics grades was identified. Their math grades and GPA, by contrast, turned out to be within the normal range suggesting domain-specific underachievement. The female physics underachievers, moreover, showed a low interest and self-concept in physics compared to the other students, complementing the picture. An independent sample was used to validate the student profiles. We finally discuss implications for physics classrooms and future research.

Keywords: *Underachievement; Gender differences; Physics education; Latent profile analysis; Science education*

Introduction

Recent reviews that summarized work on women in science identified secondary school as crucial point in time to consolidate gender differences in achievement, engagement, interest, and participation in science (see Ceci, Ginther, Kahn, & Williams, 2014; Ceci, Williams, & Barnett, 2009). These gender differences are also reflected in the smaller proportion of talented, intelligent females who specialize in science (e.g., Lubinski & Benbow, 1992). Intelligent students who fail to realize their potential especially in physics have become a growing concern in today's competitive, technology-dependent society. And in light of the current state of research, there is reason to expect more girls than boys among such physics underachievers, contributing to the gender gap in physics. While gender differences have been addressed in terms of both general scholastic underachievement (Colangelo, Kerr, Christensen, & Maxey, 1993) and general physics attainment (e.g., Heilbronner, 2012; Lubinski & Benbow, 1992, 2007), in this study we investigate gender-specific underachievement in physics (c.f. C. M. Adams, 1996; Reis, 1991).

In the present study, we want to contribute to a more precise picture of the gender-specific prevalence of physics underachievers. Profound knowledge about this student group constitutes the basis for further research and school interventions that may reduce the gender gap in physics. By using multiple group latent profile analysis, we propose an innovative statistical approach to determine physics underachievers. Student profiles were defined by a measure of intellectual potential and physics grades. The domain-specificity of physics underachievement was investigated by analyzing the underachieving students' performance in other school subjects. We examined the physics underachievers' interest and self-concept in physics to further describe this group of students.

To set the stage for this study, in the following sections we start from the broad perspective of general scholastic underachievement and increasingly zoom in on characteristics of physics underachievers leading to gender differences in physics, and, finally, to the research questions of the present study.

Operational Definitions of Underachievement

As a preliminary remark, underachievement research suffers from a similar phenomenon as its objects of study: a failure to exploit its potential. One reason for this is the fragmented research base. Definitions of underachievement vary considerably across studies. Hence, comparing results and drawing general conclusions is difficult, which has severely hampered scientific progress (see Dowdall & Colangelo, 1982; Ziegler, Ziegler, & Stoeger, 2012).

According to Reis and McCoach (2000), definitions of underachievement can be categorized in four different ways. A first approach is to determine a quantified discrepancy between a person's potential and achievement (e.g., more than one standard deviation discrepancy between the standardized ability and achievement measures). A second category subsumes studies that speak of underachievement when a person's scores exceed certain cut-off values for intellectual potential (e.g., $IQ \geq 130$) and fall below a defined level of school achievement (e.g., grade $\leq C$ in the US scales). A third way is based on regression analysis. Hence, the existence of a substantial discrepancy (e.g., more than one standard error of the regression) between actual school achievement and the one predicted by a student's intellectual potential determines underachievement in the third category (e.g., Lau & Chan, 2001). In the last category, learners are called underachievers simply if they fail to take advantage of their latent intellectual potential (see Gagné, 2004, 2005).

Educational psychologists have been studying students who underachieve for about 70 years now (e.g., Conklin, 1940; McCall, 1994; Reis & McCoach, 2000; Siegle, 2013; Thorndike, 1963). In the course of these many years, underachievement research had to take a lot of criticism. In addition to the heterogeneity of definitions (e.g., Siegle, 2013; Smith, 2003; Thorndike, 1963), critics further list a number of methodological shortcomings. For instance, when cut-off values or a certain discrepancy between potential and achievement are used to define underachievement, the measurement errors inherent in any psychological assessment are neglected. Applying these operational definitions of underachievement, Ziegler and colleagues (2012) could exemplarily show how the number of underachievers, given a certain true number, is severely overrated due to measurement errors. Moreover, by using cut-off values or a discrepancy, the at least ordinal variables intellectual potential and achievement are used to rather arbitrarily create distinct categories of normal, high, or underachievement (Reis & McCoach, 2000). In the regression analytic approach to define underachievement, the estimation of the regression is based on the whole student sample that

also encompasses the to-be-detected underachievers. Consequently, the standard error of estimation, whose magnitude is commonly used to determine underachievement, is biased because the regression itself is biased by the underachievers in the sample. To sum up, justified criticism led to a decline in studies on scholastic underachievement in recent decades. While the construct of underachievement is definitely of substantial value, it is the method that has to be reconsidered.

To avoid the common points of criticism, we decided to apply latent profile analysis (LPA) with two indicator variables measuring intellectual potential and physics achievement to operationalize underachievement. LPA is a type of mixture model that is, in highly simplified terms, estimating the existence of subgroups or profiles within an overall sample based on the similarity on certain continuous indicator variables, without requiring that neither profile sizes nor characteristics are defined before (Gibson, 1959; Lazarsfeld & Henry, 1968; Vermunt & Magidson, 2002). Hereby, the analysis seeks to explain similarity on the continuous indicator variables by relating the similarity to a newly introduced categorical latent variable that defines the profiles. The number of profiles that are estimated has to be specified by the user. As a result, the LPA produces indicators of the quality of the respective profile solution, profile sizes and characteristics as well as the profile membership probabilities for every person.

Using LPA, methodological problems that accompany cut-off values and a priori defined discrepancies can be circumvented. So LPA allows for classification uncertainty since membership in any profile is represented as probability. Thus, a student is not assigned to one distinct profile postulating that this student either is, or is not, an underachiever, for instance. Rather, there are variables created indicating profile membership probability for every profile for every student. The LPA aims at describing the whole student sample in the form of profiles and is not only geared to the categorization of students into underachievers and non-underachievers. So students are clustered based on the similarity on the intellectual potential and achievement indicator variables. This characteristic also eliminates the problem of using standard errors that are potentially biased by underachievers in the sample as decision criterion to distinguish between underachievers and non-underachievers, as it is done in the regression analytic approach. The operational definition by means of LPA provides a clear instruction of how to proceed and enables comparison and replication across studies.

Motivational Correlates of Physics Underachievement

Academic interest and self-concept are two variables that have often been associated with both school achievement and general scholastic underachievement. There is broad evidence that self-concept and school achievement influence each other, presumably in the sense of reciprocal effects (see Marsh & Craven, 2006; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005). A similar reciprocal relationship is assumed for interest and school achievement (see Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Köller, Baumert, & Schnabel, 2001; Schiefele, Krapp, & Winteler, 1992). Moreover, both academic interest and self-concept could be expected to be negatively related to boredom (Pekrun, Hall, Goetz, & Perry, 2014). Pekrun and colleagues (2014) suggest a reciprocal relationship between boredom and school achievement, too. To conclude, self-concept and interest seem to be related to school achievement in both direct and indirect ways.

In their theoretical paper, Snyder and Linnenbrink-Garcia (2013) propose several motivational factors suggested by existing research on achievement-motivation that might contribute to the development of general underachievement in gifted students on consecutive developmental stages in two postulated pathways. On the third stage, which starts with the entry in secondary school, the authors assume a decrease in academic self-concept due to the Big Fish Little Pond Effect (see e.g., Marsh, 1987) leading to coping mechanisms such as disidentification with academics and disengagement. In the alternative pathway to underachievement, students in secondary school who experience enhanced academic challenge may consider the costs (such as effort and time) of academics as increasingly high and therefore suffer from decreasing utility, intrinsic, and attainment value concerning academics, again leading to disengagement and disidentification. While the first pathway may particularly be evidenced by decreased academic self-concept, the second pathway may especially be reflected in decreased interest in academics.

In line with these considerations, literature reviews reported a poor academic self-concept as frequent characteristic of general scholastic underachievers (McCall, 1994; Reis & McCoach, 2000). According to Sparfeldt, Schilling, and Rost (2006) gifted underachievers showed strikingly lower scores on an academic self-concept scale in primary school than their achieving peers. Referring to the domain of physics, top performers in science are characterized by an exceptionally high self-concept in science (Organisation for Economic Co-operation and Development, 2009). Therefore, low self-concept in physics may

accompany physics underachievement. Moreover, gifted high achievers and gifted underachievers were also found to differ regarding their interest in classes, with more positive attitudes on the part of the high achievers (McCoach & Siegle, 2003). In the domain of physics, excellent achievement turned out to be associated with particularly high interest in physics or science in general (W. K. Adams et al., 2006; Lubinski & Benbow, 2006; Organisation for Economic Co-operation and Development, 2009; Robertson, Smeets, Lubinski, & Benbow, 2010). Consequently, low interest in physics can be expected to accompany physics underachievement, too.

There are more motivational variables that can be assumed to be associated with physics underachievement. Yet, as summarized above, the interplay between academic interest and self-concept, on the one hand, and academic achievement and underachievement, on the other hand, is evidenced by extensive research. Linking our knowledge from this field of research to findings in the domain of physics, existing research suggests a deficit on the part of physics underachievers in terms of interest and self-concept in physics.

Gender Differences and Physics Underachievement

Females have been quite consistently found to achieve on a lower level in physics than males (e.g., Beaton et al., 1996; Lubinski & Benbow, 1992; Organisation for Economic Co-operation and Development, 2009; Taasooobshirazi & Carr, 2008). In line with this, male students seem to be equipped with a higher self-concept in physics than female students (e.g., Debacker & Nelson, 2000; Heilbronner, 2012; Hoffmann, 2002; Jansen, Schroeders, & Lüdtke, 2014; Schober, Reimann, & Wagner, 2004). There is also broad evidence that girls are less interested in physics than boys (e.g., W. K. Adams et al., 2006; Hart, 1996; Hoffmann, 2002; Kahle & Lakes, 1983; Murphy & Whitelegg, 2006b). This general finding is closely related to the result that males seem to be more interested in things whereas females appear to show higher interest in people (Su, Rounds, & Armstrong, 2009). Seidel (2006) used latent class analysis to divide ninth-grade Gymnasium physics students into five classes based on their general cognitive ability, physics knowledge (standardized test items), interest in physics, and physics self-concept. One of the resulting classes was called the “uninterested” students (high on general cognitive ability, mixed on physics knowledge, very low on interest, and intermediate self-concept) and another class was labeled the

“underestimating” students (high on general cognitive ability and physics knowledge, intermediate interest, and very low self-concept). Importantly, a considerably higher proportion of girls than boys belonged to these profiles.

Although, overall, girls outperform boys in school (e.g., Voyer & Voyer, 2014), the girls’ advantage seems to disappear in physics classes (e.g., Deary, Strand, Smith, & Fernandes, 2007). Taken together, there is good reason to expect at least some female underachievers in the domain of physics.

The Present Study

The present study aimed at investigating underachievement in physics in terms of its gender-specific prevalence expecting especially girls to be physics underachievers. Hence, we hypothesized to find a higher proportion of girls than boys among physics underachievers defined by means of latent profile analysis.

The domain-specificity of physics underachievement was examined by analyzing the underachieving students’ performance in other school subjects. Therefore, we additionally considered the students’ corresponding GPA and in particular their math achievement which has many overlaps with physics and is also part of the STEM (science, technology, engineering, and mathematics) field. We further investigated the students’ interest and self-concept in physics expecting physics underachievers to show lower manifestations on these variables than the high achieving students.

Method

Participants

In this study, we used two existing samples to be able to investigate physics underachievement based on a sufficiently large sample size. Yet, after the student profiles had been determined in the whole student sample, the two samples were examined separately to confirm that the two samples were comparable and that the detected profile solution also fitted the data in each of the two samples.

In sample 1 that included $N = 133$ students (78 girls; age $M = 16.56$ years, $SD = 1.36$ years), students were recruited individually by advertising and contacting teachers and student representations. Sample 1 thus consisted of individual students coming from different classes and schools. Data was only gathered from German-speaking Swiss Gymnasium students from the upper secondary level who already had physics instruction. At the German-speaking Swiss Gymnasium, all students have to attend physics classes at the upper secondary level. Therefore, we did not investigate a selected population of students having explicitly opted for physics classes but the whole range of upper secondary level Gymnasium students.

Sample 2 ($N = 183$; 104 girls; age $M = 16.02$ years, $SD = 0.84$ years), by contrast, consisted of 14 upper secondary level physics classrooms from five Swiss Gymnasien that were recruited in the context of a school intervention study not discussed here (for more details, see Hofer, Stern, Rubin, & Schumacher, 2015). In all 14 classrooms, the same introductory Newtonian mechanics topics were taught during 18 physics lessons. The intervention that was compared to regular instruction did neither affect the intellectual potential indicator variable that was elicited before or after the intervention nor the physics achievement indicator variable (i.e., physics grades) that reflected the students' physics school achievement during the intervention. Interest and self-concept in physics, which were measured using the same scales as in sample 1, were assessed immediately before the intervention.

The total sample consequently comprised $N = 316$ students (182 girls; age $M = 16.25$ years, $SD = 1.12$ years). Importantly, unless otherwise specified, analyses were based on the total sample of $N = 316$ Swiss upper secondary level Gymnasium students.

Using Latent Profile Analysis to Operationalize Underachievement

In the context of latent profile analysis (LPA), a systematic co-occurrence of high scores on the intellectual potential indicator variable and low scores on the physics achievement indicator variable determined a profile of physics underachievers. Importantly, the LPA can only find what is systematically occurring in the data. The findings yielded in this explorative way have to be validated in an independent sample.

Variables and Design

In a correlational design, we applied a multiple group LPA on the student data, with the two indicator variables *intellectual potential* and *physics achievement*, to detect a stable pattern of systematically occurring student profiles. The multiple groups were defined by the known-class variable *gender*. Hence, student profiles were estimated for female and male students. In case a resulting student profile showed high intellectual potential and low physics achievement, we spoke of *physics underachievement*.

All other variables, *math grades*, *GPA*, *interest in physics*, and *self-concept in physics*, were not used to define the student profiles but to describe the educed profiles afterwards. Thus, a potential underachievers profile was further compared to all of the other student profiles in terms of the *math grades average* and *GPA average* within each of the student profiles. The student profiles were further compared in terms of the motivational variables *interest in physics* and *self-concept in physics* that were again averaged within each of the student profiles.

Procedure and Measures

The tests for measuring the students' intellectual potential were administered by two trained and experienced professionals in group testing sessions. The students' grades in physics as well as in other subjects were recorded from the students' two most recent report cards or provided by the teachers. Data regarding motivational and demographic (e.g., student gender) variables were gathered by means of the same online-survey in sample 1 and sample 2. The students received the link to the approximately 20-minutes online-survey in the course of the correspondence. The scales that were used to measure the motivational variables were adapted to physics instruction.

Intellectual potential. In the context of the data collection for a larger research project, sample 1 was presented with the "Berliner Intelligenzstruktur-Test" (BIS Test; Jäger, Süß, & Beauducel, 1997). This test enables a broad assessment of the operational abilities processing capacity, creativity, memory retention, and speed of operation and the content based abilities verbal reasoning, numerical reasoning, and figural/spatial reasoning. Reliability, validity, and objectivity are ascertained and convincing and group testing is feasible. To operationalize

intellectual potential in the present study, however, we preferred measuring general reasoning ability over measuring diverse cognitive facets. Referring to several empirical evaluations of the test's validity (Jäger et al., 1997), the scale measuring processing capacity turned out to be highly related not only to cognitive abilities such as relational reasoning, storing, and processing but also to science grades. In addition to processing capacity, also figural ability seems to be strongly associated with general reasoning ability (Bucik & Neubauer, 1996). High spatial ability, moreover, has been found to be especially important for STEM achievement (see e.g., Lubinski & Benbow, 2006; Robertson et al., 2010; Wai, Lubinski, & Benbow, 2009). Consequently, we decided to use the composite score of the five figural/spatial processing capacity problems (such as solving analogies or continuing logical progressions) to estimate students' intellectual potential in this study. The scale's Cronbach's α was satisfactory ($\alpha = .70$).

In sample 2, the students' intellectual potential was measured by means of the well-established set II score of Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1992) that explicitly assesses reasoning ability. Hence, the measures of both samples could be considered to be indicators of the basic cognitive ability of general reasoning representing the students' intellectual potential in this study. To obtain one measure that could be used as estimate for the students' intellectual potential in the total sample, the score of the figural/spatial processing capacity in sample 1 and the set II score in sample 2 were z-standardized. The resulting joint scale is simply referred to as *intellectual potential* in the following.

Physics achievement, math achievement, and GPA. Physics achievement, math achievement, and GPA were assessed by means of grades. We used grades and not standardized achievement tests, since that is what the students get as feedback at school, to what they react, and what considerably influences further school engagement (Poorthuis et al., 2014; von Maurice, Dörfler, & Artelt, 2014). While math was intentionally considered separately (due to its conceptual proximity to physics), Biology, German, and English grades were averaged to obtain a GPA that was used in all analyses. Whenever available, grades from two report cards were used to calculate an average for each subject. In Switzerland grades range from 6 to 1 with smaller numbers indicating lower performance. With grades lower than 4 students fail.

Interest in physics. The scale to measure interest in physics was adopted from the international student-survey of PISA 2006 (Frey et al., 2009). It consists of four items with four-point Likert scales spanning from 0 “completely disagree” to 3 “completely agree” (Cronbach’s $\alpha = .87$; sample item: “These days I like dealing with physics problems.”; see Appendix A).

Self-concept in physics. The students’ self-concept in physics was elicited adapting four items of the “DISK-Gitter mit SKSLF-8” (Rost, Sparfeldt, & Schilling, 2007), a published test in German language targeting school subject specific self-concept. Students can choose between six answer alternatives spanning from 0 “not true for me at all” to 5 “exactly true for me” (sample item: “These days I feel that I can solve problems in physics easily.”; see Appendix A). The reliability of the scale was high with Cronbach’s $\alpha = .94$.

Data Analysis

The latent profile analyses were run with the software Mplus Version 7.11 (Muthén & Muthén, 2012). We applied robust maximum likelihood estimation to potentially correct chi-square based fit statistics and all parameter estimates’ standard errors for leptokurtic or platykurtic data. The z-standardized intellectual potential and physics grades average were used as indicator variables. To investigate gender-specific physics underachievement, gender was included as known-class variable. This means that latent profiles were estimated for girls and for boys in the form of a multiple group LPA or mixture model (see Muthén & Muthén, 2012). A six-step procedure was chosen to examine gender differences in physics underachievement, to consider the students’ school achievement in subjects other than physics, and, finally, to inspect the students’ interest and self-concept in physics. Accordingly, first of all, the number of profiles had to be identified, second, gender differences in the profile formation were examined, third, the resulting profiles were validated in each of the two samples, fourth, the resulting profiles were validated in an independent validation sample, fifth, differences between the profiles regarding math achievement and GPA were investigated, and, sixth, differences between the profiles regarding students’ interest and self-concept in physics were investigated.

Step 1: Identification of profile number. In a first step, the number of profiles was determined. Models with two to 14 profiles, with the known-class variable gender, were

realized. Expecting more than 14 gender-specific profiles was considered practically and theoretically unreasonable. The analysis was allowed to compute the profiles without any restrictions with regard to the gender variable (i.e., independently for girls and boys), because the model should not be constrained before the best-fitting number of profiles was determined. Hence, the same number of profiles was estimated independently for female and male students. Consequently, only profile solutions with an even number of profiles were realized in this step. Model-fit then was compared between the profile solutions. It is not possible to perform significance tests for general model-fit in mixture models with known-classes, since there is no unrestricted model to test against. The model-fit, the correspondence between data and the specified latent profile model, was therefore primarily evaluated by inspecting Information Criteria (IC) looking for the solutions with the lowest (i.e., best) IC values (see e.g., Geiser, 2011; Gollwitzer, 2012). The ICs take account of the model's logarithmized Likelihood ($\text{Log } L$), the number of model parameters (k), and, for most criteria, also the sample size. There is no definite answer to the question which IC to use in the context of latent class or profile analyses. However, simulation studies recommend an examination of the sample-size adjusted Bayesian Information Criterion (aBIC; Sclove, 1987) or of the standard Bayesian Information Criterion (BIC; Schwarz, 1978), tending to favor the former (Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2007; Yang, 2006; Yang & Yang, 2007). Hence, in the first place the aBIC and in the second place the BIC were considered to assess model-fit with relatively lower values indicating better model-fit.

Step 2: Gender differences. In a second step, the profile solution determined in step one was used to investigate whether girls and boys differ. By testing whether the profile-specific estimates of the indicator variables are comparable between girls and boys, measurement invariance is tested in the context of LPA (c.f. Eid, Langeheine, & Diener, 2003; Specht, Luhmann, & Geiser, 2014). Different models which realized different degrees of conformity between girls and boys were compared. The least restrictive model suggested unique profiles for girls and boys with all of the profile-specific indicator variable means estimated freely. Then, successively, the most similar profiles were constrained to be equal (i.e., indicator variable means estimated at once) between girls and boys. The nested models' fit to the data was contrasted to find the model which best described the degree of conformity between females' and males' profiles in the sample. Hence, models were compared using the aBIC and log-likelihood tests. Log-likelihood tests compare more restrictive models with less

restrictive but nested models via a chi-square distributed test statistic that yields a p -value (for detailed information on the test, see UCLA: Statistical Consulting Group, 2014). So, significant change in the log-likelihood values (adjusted by scaling correction factors), relative to change in the associated degrees of freedom, was used to determine whether increasingly restrictive nested models can be warranted. If there were no significant discrepancies in model-fit, the most restrictive model was chosen.

Step 3: Profile validation across samples. To achieve a sufficiently large sample size, we used two samples to investigate underachievement. Although the two samples were highly comparable regarding the participating students and the measures implemented, they nevertheless differed in the operationalization of the intellectual potential measure and the recruitment of the students. Accordingly, we checked that the solutions obtained separately within each of the two samples did not differ significantly from each other and from the final solution based on the total sample. The fact that we had to combine two samples thus allowed us to control for method variance and validate that the final student profiles were theoretically meaningful across methodological variations.

Hence, first, we aimed to show that the profile solution obtained independently with sample 1 did not differ from the profile solution obtained with sample 1 when the indicator variables, however, were fixed at the profile-specific estimates from the final profile solution obtained with the total sample in step two. Therefore, in the latter model, the final profile solution that was obtained with the total sample was imposed on sample 1. The fit of the two nested models was compared using the aBIC and the log-likelihood test. In case of no significant differences between the fit of the two nested models, we could assume the final profile solution to hold true for sample 1, too. An analogue analysis was conducted for sample 2.

Second, we aimed to show that the solution obtained independently with sample 1 did not differ significantly from the solution obtained independently with sample 2. To test this assumption, the indicator variable estimates, resulting from the independent profile estimation in sample 2, were used as default values for the indicator variables of the profile estimation in sample 1. The corresponding profile solution that hence reflected the solution obtained independently with sample 2 imposed on sample 1 was compared to the solution obtained independently with sample 1. Again, the fit of the two nested models was compared using the aBIC and the log-likelihood test. No significant discrepancies in this case would

indicate that the independent profile solutions in the two samples did not differ significantly and that both samples hence warranted the final profile solution.

Step 4: Validation of the profiles. We further tried to validate the final student profiles in an independent validation sample. The validation sample comprised $N = 264$ (143 female students) German-speaking Swiss Gymnasium students from the upper secondary level with a mean age of $M = 15.32$ years ($SD = 1.15$ years) who already had physics instruction. Physics achievement was measured using the students' performance in one physics examination and intellectual potential was again assessed by means of the set II score of Raven's Advanced Progressive Matrices. Inspired by the approach described by Finch and Bronk (2011) in the context of confirmatory latent class analysis, the indicator variables in the validation sample were fixed at the profile-specific estimates from the final profile solution. The resulting profile structure was interpreted and the model fit of this restrictive model was compared with the model fit of the corresponding unrestrictive model.

Step 5: Differences in math achievement and GPA. The model that resulted as the final model from the first two steps was used in the steps five and six. In order to find out more about the domain-specificity of physics underachievement, math grades and GPA were directly included in the final model as external outcome variables. In doing so, the probabilistic nature of profile membership was taken into account. Consequently, the analysis was less afflicted with disregarded errors than an independent analysis that deterministically categorizes students based on a most likely latent profile membership variable (c.f. Asparouhov & Muthén, 2012). While the two indicator variables (intellectual potential and physics grades average) defined the student profiles, external outcome variables should further describe the profiles but not affect the profile estimation. Therefore, the manual 3-step approach as described by Asparouhov and Muthén (2012) was conducted to separately include and estimate each external outcome variable. To examine whether the estimated means of the external outcome variables significantly differed between underachievers and other student profiles, the chi-square value (χ^2) of the Wald test of parameter constraints was used to test the null hypothesis of parameter equality between two student profiles with always one degree of freedom owing to the pairwise comparisons ($df = 1$).

Step 6: Differences in interest and self-concept. Following the same approach as described in step five, now the motivational variables interest and self-concept were each directly included in the final model as external outcome variables.

Results

Descriptive Statistics

Descriptive statistics and intercorrelations between intellectual potential, physics achievement, math achievement, GPA, interest in physics, and self-concept in physics can be found in Table 2.1.

Table 2.1

Intercorrelations, Means, Standard Deviations, and Scale of the Measures Used in this Study

Measure	1	2	3	4	5	<i>M</i>	<i>SD</i>	<i>Scale</i>
1. Intellectual potential ^a	-					515.07 27.35	32.84 4.19	389-611 0-36
2. Physics achievement	.11	-				4.61	0.64	1-6
3. Math achievement	.31**	.51**	-			4.54	0.74	1-6
4. GPA	.04	.32**	.35**	-		4.63	0.43	1-6
5. Interest in physics	.18**	.38**	.28**	.14*	-	1.40	0.71	0-3
6. Self-concept in physics	.16**	.48**	.33**	.11	.69**	2.43	1.16	0-5

Note. * $p < .05$. ** $p < .01$.

^a due to z-standardization, this measure's mean $M = 0.00$ and standard deviation $SD = 1.00$. To provide more informative statistics, means and standard deviations of the figural/spatial processing capacity (for sample 1) and of the set II score of Raven's Advanced Progressive Matrices (for sample 2) are reported.

Step 1: Identification of Profile Number

In the comparison of the models with two to 14 profiles, the six-profiles-solution turned out to fit the data best (see Table 2.2). The six-profiles-solution showed the lowest (i.e., best)

aBIC. The BIC that sanctions complexity more than other information criteria, especially in larger samples, consistently increased with an increasing number of profiles and thus would have recommended a two-profiles-solution (see Bacci, Pandolfi, & Pennoni, 2014). Discussing the problem of deciding about the goodness of latent profile (or class) models, Marsh, Lüdtke, Trautwein, and Morin (2009) recommended that a solution should be theoretically meaningful, parsimonious, as well as interpretable. In line with this idea, the two-profiles-solution, as suggested by the BIC, would be of limited value in light of the underlying theoretical considerations on gender-specific underachievement. Consequently, the six-profiles-solution was chosen to proceed.

Table 2.2

Logarithmized Likelihood (Log L), Number of Parameters (k), aBIC, and BIC for Different LPA Solutions with Known-Class Gender

Number of profiles	Log L	k	aBIC	BIC
2	-1099.92	7	2217.94	2240.14
4	-1091.39	13	2216.37	2257.60
6	-1078.30	19	2205.69	2265.95
8	-1071.90	25	2208.40	2287.69
10	-1066.44	31	2212.99	2311.32
12	-1059.55	37	2214.71	2332.07
14	-1055.01	43	2221.13	2357.51

Note. Values in bold typeface indicate the profile solution favored by the respective criteria.

Step 2: Gender Differences

Investigating measurement invariance. After the number of profiles was determined resulting in a six-profiles-solution, measurement invariance between girls and boys was examined. The model for which the profile with the highest gender similarity was constrained to be invariant between girls and boys fitted the data not significantly worse ($LL\ p = .64$) than the unconstrained model where all profiles were estimated freely for girls and boys.

Moreover, the aBIC was lower for the more restrictive model (aBIC = 2201.25) than for the unconstrained model (aBIC = 2205.69). An additional profile, with the second highest gender similarity, was constrained to be equal between female and male students. This even more restrictive model, however, proved to fit the data significantly worse than the model with only one constrained profile ($LL\ p < .01$) and also the aBIC slightly increased again (aBIC = 2202.31). Hence, profiles differed between females and males with the exception of only one profile that showed measurement invariance in terms of gender. Consequently, five distinct student profiles emerged. The five-profiles-solution thus resulted as the final solution. An inspection of the residual statistics further indicated that the final solution yielded a good approximation of the empirical means and variance/covariance structures. Figure 2.1 depicts the five profiles with the mean estimated scores on the two indicator variables (z-standardized intellectual potential and physics grades average) for female and male students.

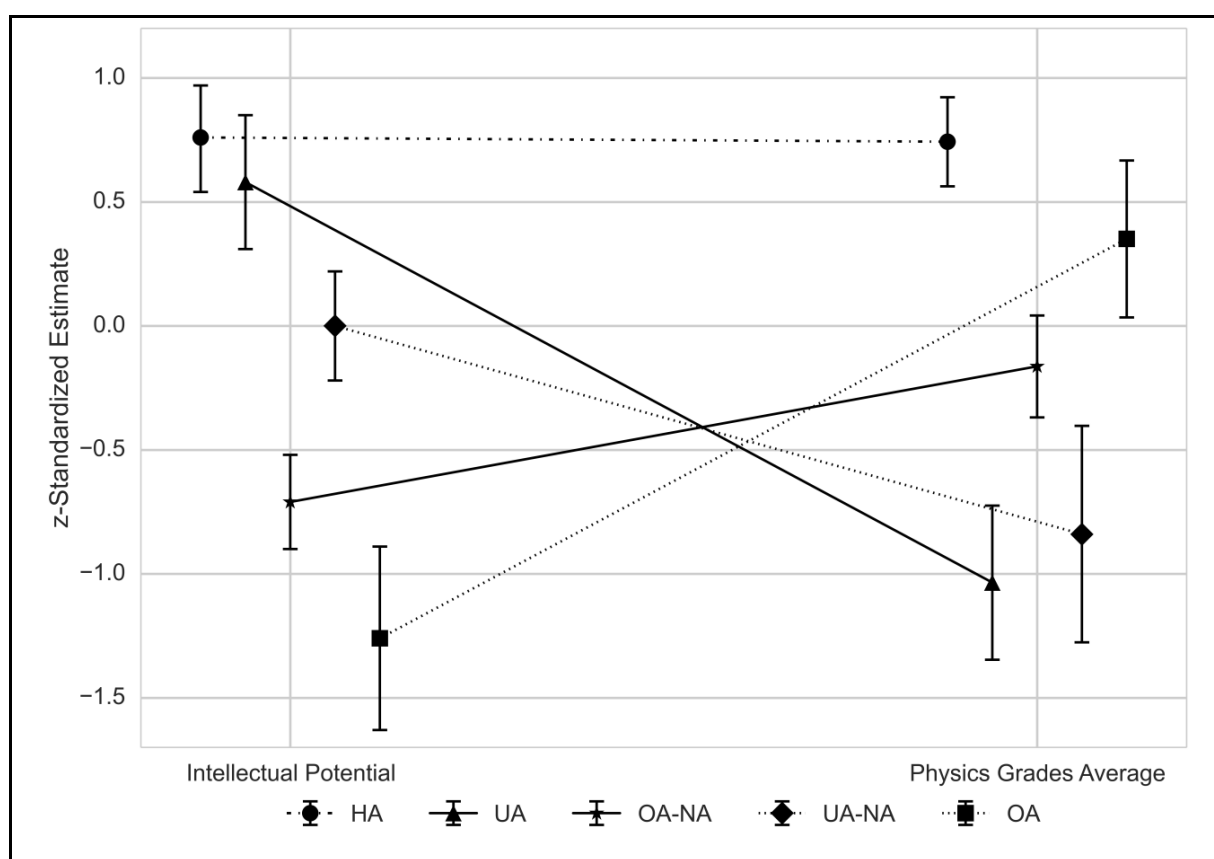


Figure 2.1. All five student profiles based on the final model. HA = high achievers, UA = underachievers, OA-NA = over-to-normal achievers, UA-NA = under-to-normal achievers, OA = overachievers. A continuous line represents females, a dotted line represents males, and the dash-dotted line represents the gender-invariant profile. Error bars represent the 95% confidence intervals.

Description of girls' and boys' profiles: Detecting underachievers. Although this article focuses on underachievers, it is necessary to describe the whole picture and briefly interpret each of the student profiles. At first, however, it is important to note that the student profiles were not classroom-dependent in sample 2, where whole classrooms participated. Hence, particular student profiles were not accumulated in single classrooms, as indicated by the most likely latent profile membership of each student, but always turned out to be distributed over at least eight of the 14 classrooms. Moreover, it has to be considered that all group size specifications provided for each of the student profiles in the following section were based on the most likely latent profile membership patterns, although, effectively, a student was not deterministically but probabilistically assigned to each of the profiles.

The profile that was invariant between females and males was the high achievers profile (16% of all girls and 51% of all boys) with very high z-standardized intellectual potential ($M = 0.76$) and z-standardized physics grades ($M = 0.74$) for both genders. In addition to the high achievers profile, within the girls, there was a clear underachievers profile (29% of all girls) with high intellectual potential ($M = 0.58$) and very low physics grades ($M = -1.04$) and an over-to-normal achievers profile with rather low intellectual potential ($M = -0.71$) and average physics grades ($M = -0.16$).

In addition to the high achievers profile, within the boys, there was an under-to-normal achievers profile with average intellectual potential ($M = 0.00$) and rather low physics grades ($M = -0.84$) and an overachievers profile with very low intellectual potential ($M = -1.26$) and high physics grades ($M = 0.35$).

To conclude, the LPA detected a profile of physics underachievers, but only among the female students. For information about profile membership proportions and counts as well as the profile-specific estimated intellectual potential and unstandardized physics grades average, see Table 2.3.

Table 2.3

Profile-Specific Membership Proportions and Counts (N) Based on the Most Likely Latent Profile Membership Pattern as well as Estimated Intellectual Potential and Unstandardized Physics Grades Average

Student profile		% of same gender	N	Intellectual potential [95% CI]	Physics grades average [95% CI]
High achievers		16.48 (girls) 50.75 (boys)	98	0.76 [0.54, 0.97]	5.09 [4.97, 5.20]
Female	Underachievers	29.12	53	0.58 [0.31, 0.85]	3.95 [3.75, 4.15]
	Over-to-normal achievers	54.40	99	-0.71 [-0.90, -0.52]	4.51 [4.37, 4.64]
Male	Under-to-normal achievers	22.39	30	0.00 [-0.22, 0.22]	4.07 [3.79, 4.35]
	Overachievers	26.87	36	-1.26 [-1.63, -0.89]	4.84 [4.63, 5.04]

Note. CI = confidence interval. Grades in Switzerland range from 6 to 1 with 6 indicating the best and 1 the worst grade. Since variance was constrained to be equal across profiles, the overall standard deviation was $SD = 0.65$ for intellectual potential and $SD = 0.48$ for the unstandardized physics grades average.

Step 3: Profile Validation Across Samples

Both according to aBIC values and the log-likelihood test ($LL\ p = .70$), the profile solution that was based on the final profile solution imposed on sample 1 did not fit the data significantly worse than the profile solution obtained independently in sample 1. The same was true regarding sample 2 ($LL\ p = .86$). Moreover, the profile solution that reflected the solution obtained independently in sample 2 did not fit the data in sample 1 significantly worse than the profile solution obtained independently in sample 1 ($LL\ p = .12$). The independent profile solutions in the two samples were comparable with each other and with the final profile solution as obtained in the total sample suggesting that the differences in the operationalization of the intellectual potential measure and in the recruitment of the students

did not affect the profile structure. These results hence validated the use of the total sample and also indicated the profiles' theoretical meaningfulness across methodological variations.

Step 4: Validation of the Profiles

We used the validation sample to validate the final profiles. Because the final profile solution fitted the data in the validation sample significantly worse than the unrestrictive solution, we relaxed the restrictions. Now only the intellectual potential indicator variable was set at the values of the final five-profiles-solution to predetermine the general structure. The physics achievement indicator variable was estimated freely to allow some variance. The resulting solution did not fit the data significantly worse ($LL\ p = .14$) than the unrestrictive solution. The aBIC favored the more restrictive model, too (1818.57 vs. 1823.85). The aBIC value of the more restrictive model was also smaller than the aBIC values of the models with two to 14 profiles. With the only exception of the profile of female over-to-normal achievers, all student profiles could be confirmed. Instead of the female over-to-normal achievers, a profile of female low achievers with rather low manifestations on both indicator variables was detected. In the validation sample, however, not the final record card physics grades but only the performance in one physics examination was available to measure physics achievement. This measure has to be considered less valid than record card grades to define underachievement because it reflects only the objective performance in one written physics test. Teachers' evaluations representing a student's performance during a whole term may differ to some extent from a student's performance in one written examination resulting in slightly distorted profiles. Although four of the student profiles, including the female physics underachievers, could be replicated, additional confirmatory analyses using record card physics grades as physics achievement indicator variable are needed to further validate the student profiles educed in the present study.

Step 5: Differences in Math Achievement and GPA

In order to find out more about the domain-specificity of physics underachievement, math grades and GPA were included in the final model as external outcome variables. For a more differentiated picture, the external outcome variables of female and male high achievers were examined separately instead of considering one joint high achievers profile. The results

regarding the comparison of the underachieving girls with all of the other student profiles are reported in more detail in the following. For all comparisons, the Wald test of parameter constraints was used. Figure 2.2 depicts the profile-specific estimated math grades and GPA as compared to physics grades and illustrates, *inter alia*, that the underachieving females' physics grades were on a considerably lower level than their math grades and GPA.

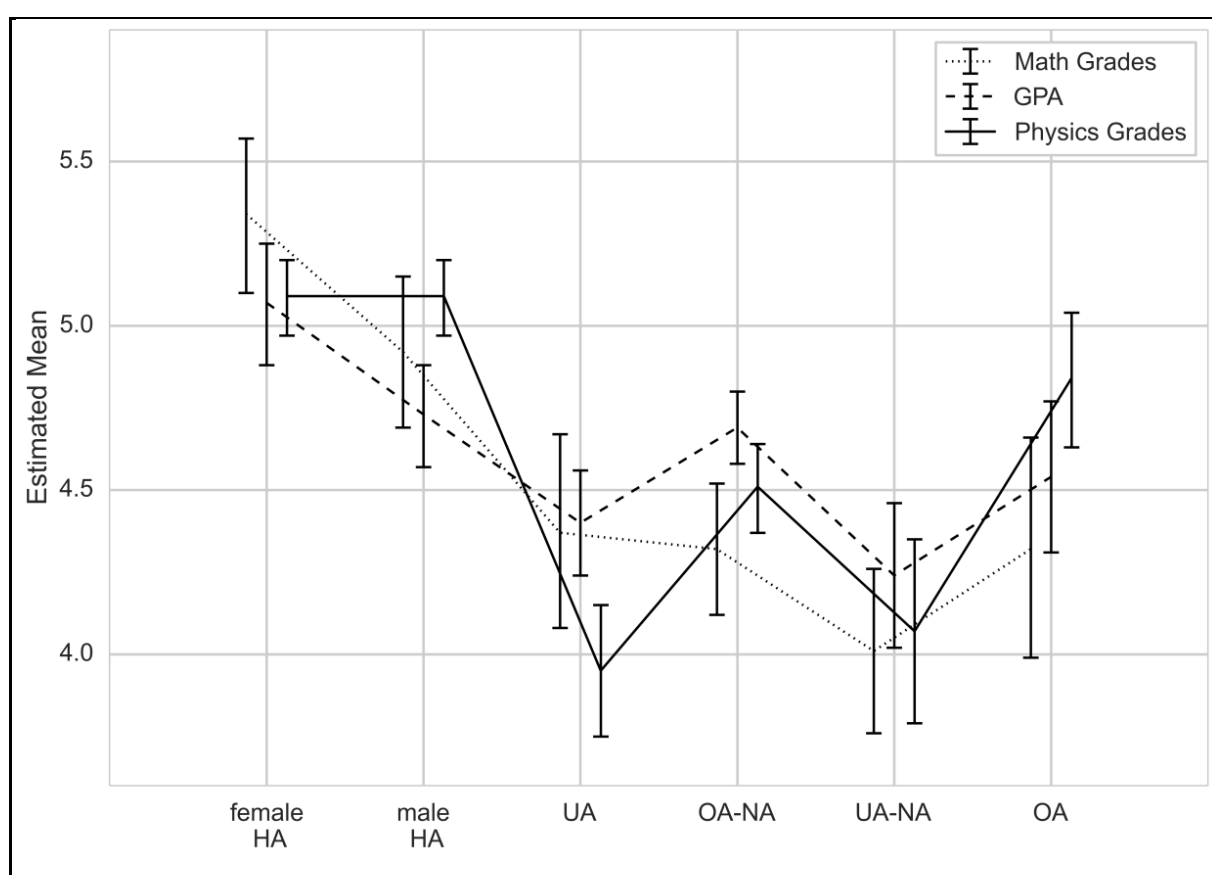


Figure 2.2. Profile-specific estimated math grades, GPA, and physics grades. HA = high achievers, UA = female underachievers, OA-NA = female over-to-normal achievers, UA-NA = male under-to-normal achievers, OA = male overachievers. Error bars represent the 95% confidence intervals. Grades in Switzerland range from 6 to 1 with 6 indicating the best and 1 the worst grade.

Regarding math grades, the underachieving girls showed significantly lower grades ($M = 4.37$) than both the female and male high achievers ($M = 5.34$, $\chi^2 = 28.90$, $p < .001$ and $M = 4.92$, $\chi^2 = 8.04$, $p < .01$), however these two groups also significantly outperformed all of the other student profiles. The underachieving girls did not differ significantly in terms of their

math grades from all of the other student profiles (all χ^2 s ≤ 0.07 , all $ps \geq .80$) and they even considerably exceeded the male under-to-normal achievers ($M = 4.01$, $\chi^2 = 3.37$, $p = .07$).

The underachieving female students' GPA ($M = 4.40$) was significantly lower than the female high achievers' GPA ($M = 5.07$, $\chi^2 = 28.74$, $p < .001$) and the male high achievers' GPA ($M = 4.73$, $\chi^2 = 8.60$, $p < .01$). Moreover, the underachieving females' GPA differed significantly from the female over-to-normal achievers' GPA, too ($M = 4.69$, $\chi^2 = 8.06$, $p < .01$). Yet, their GPA was not significantly lower than the GPA of both the male under-to-normal achievers and the male overachievers (all χ^2 s ≤ 1.34 , all $ps \geq .25$). Table 2.4 provides the estimated means, standard deviations, and 95% confidence intervals for math grades and GPA for all of the student profiles.

Step 6: Differences in Interest and Self-Concept

Table 2.4 also lists the estimated means, standard deviations, and 95% confidence intervals for interest and self-concept in physics for all of the student profiles. Again, female and male high achievers were considered separately and the Wald test was used for all comparisons. The underachieving girls showed the lowest interest in physics of all of the student profiles ($M = 1.02$). They significantly differed from the female and male high achievers ($M = 1.74$ and $M = 1.91$) as well as the male overachievers ($M = 1.49$, all χ^2 s ≥ 5.18 , all $ps < .05$). They did not differ significantly from female over-to-normal achievers and male under-to-normal achievers (all χ^2 s ≤ 1.93 , all $ps \geq .16$).

The same pattern resulted for the students' self-concept in physics. The underachieving females had the lowest self-concept in physics of all of the student profiles ($M = 1.69$). They significantly differed from the female and male high achievers ($M = 2.94$ and $M = 3.45$) as well as the male overachievers ($M = 2.68$, all χ^2 s ≥ 11.19 , all $ps < .001$). They did not differ significantly from female over-to-normal achievers and male under-to-normal achievers (all χ^2 s ≤ 2.34 , all $ps \geq .13$).

Table 2.4

Profile-Specific Estimated Means, Standard Deviations, and 95% Confidence Intervals for Math Grades, GPA, Interest, and Self-Concept

External outcome variable	Student profile		<i>M</i>	<i>SD</i>	95% CI
Math grades (scale 1-6)	Female	High achievers	5.34	0.49	[5.10, 5.57]
		Male High achievers	4.92	0.60	[4.69, 5.15]
	Female	Underachievers	4.37	0.61	[4.08, 4.67]
		Over-to-normal achievers	4.32	0.66	[4.12, 4.52]
	Male	Under-to-normal achievers	4.01	0.51	[3.76, 4.26]
		Overachievers	4.32	0.77	[3.99, 4.66]
GPA (scale 1-6)	Female	High achievers	5.07	0.37	[4.88, 5.25]
		Male High achievers	4.73	0.35	[4.57, 4.88]
	Female	Underachievers	4.40	0.35	[4.24, 4.56]
		Over-to-normal achievers	4.69	0.37	[4.58, 4.80]
	Male	Under-to-normal achievers	4.24	0.41	[4.02, 4.46]
		Overachievers	4.54	0.34	[4.31, 4.77]
Interest in physics (scale 0-3)	Female	High achievers	1.74	0.50	[1.52, 2.97]
		Male High achievers	1.91	0.65	[1.69, 2.12]
	Female	Underachievers	1.02	0.60	[0.77, 1.28]
		Over-to-normal achievers	1.11	0.59	[0.96, 1.26]
	Male	Under-to-normal achievers	1.30	0.58	[1.00, 1.61]
		Overachievers	1.49	0.75	[1.18, 1.80]
Self-concept in physics (scale 0-5)	Female	High achievers	2.94	0.71	[2.62, 3.27]
		Male High achievers	3.45	0.70	[3.23, 3.67]
	Female	Underachievers	1.69	0.90	[1.35, 2.03]
		Over-to-normal achievers	1.91	1.12	[1.64, 2.18]
	Male	Under-to-normal achievers	2.14	0.94	[1.68, 2.59]
		Overachievers	2.68	1.13	[2.21, 3.15]

Note. CI = confidence interval.

Discussion

This study drew on the construct of underachievement to shed light on unexpectedly low performance of able students in Gymnasium physics instruction, focusing especially on gender differences. With an operational definition of underachievement based on latent profile analysis, underachievers in the sample were successfully detected, described, and validated in an independent sample. A five-profiles-solution with only one of the profiles, the physics high achievers, showing measurement invariance across gender turned out to fit the data best. Among these five student profiles, physics underachievers existed only for female but not for male students, clearly confirming prior expectations. The underachieving girls, representing 29% of all females in the total sample, showed average school performance with regard to GPA and math grades. The latter is of particular importance, taking into account the many overlaps between the two STEM subjects math and physics. The findings suggest that the girls' underachievement is especially prominent in the domain of physics, although they seem to not fully exploit their high intellectual potential in other school subjects as well. Because of the severity of their underachievement in physics, we decided to nevertheless speak of domain-specific underachievement. Showing only average school performance is not especially harmful and problematic and may reflect the importance these intelligent girls currently attach to school. Their extremely low physics grades ($M = 3.95$), on the contrary, may hamper their academic careers and severely restrict future opportunities.

Male students (the male under-to-normal achievers) displayed only slight underachievement based on only average intellectual potential that could not be clearly differentiated from normal achievement. Interestingly, while real underachievement in physics appeared only for girls, real overachievement in physics appeared only for boys. Hence, some boys (the male overachievers) managed to get considerably better grades in physics than their relatively low intellectual potential would suggest. In line with the results of Hofer (2015), the findings of the present study may partly reflect a gender bias in physics teachers' grading favoring male students. Closely related, the only student profile with a high correspondence between intellectual potential and physics grades was the high achievers profile. This finding is consistent with the generally low correlation between intellectual potential and physics grades in the overall sample ($r = .11$). Although, conceptually, grades should always be used to define underachievement in the first place, it remains to be clarified how the student profiles would perform on an alternative measure of physics knowledge that

is distinct from grades and represents a more objective measure of performance. From the students' perspective, however, grades definitely have to be considered the more apparent and thus more relevant achievement measure, as evidenced by their substantial impact on academic interests, self-concepts, and future school-engagement as well as on the students' later academic and career opportunities (see Harackiewicz et al., 2008; Marsh et al., 2005; Poorthuis et al., 2014; von Maurice et al., 2014).

In line with prior expectations, physics underachievers were characterized by a considerably low interest and self-concept in physics. They showed significantly lower manifestations on these two variables than the high and overachieving students. We elaborate on this finding in the next section that derives implications for physics classrooms.

Implications for Classroom Practice

The underachieving girls seem to struggle specifically with physics classes. They displayed a very low interest in physics. Perhaps the often uninspiring physics instruction that is mainly based on memorizing and practicing formulae application with little room for deeper thinking processes (Langer Tesfaye & White, 2012; Seidel et al., 2006; Taconis, 1995) does not appeal to those intelligent girls (see e.g., Kang & Wallace, 2005; Taconis, Ferguson-Hessler, & Broekkamp, 2001; Zohar, 2006; Zohar & Sela, 2003). According to a study by Kahle and Lakes (1983), particularly 17-year old girls reported that they experience science classes as boring memorizing of facts (see also Hart, 1996). In line with this, enjoyment and interest in physics seem to diminish through schooling and this decline is reported to be more pronounced for girls than for boys (Murphy & Whitelegg, 2006a). The girls who underachieved in physics also felt rather unable to do physics and viewed themselves as performing less well than their classmates. Reviewing work on gender differences in science, Taasoobshirazi and Carr (2008) accordingly concluded that teachers evaluate the performance and capability of the girls in physics classes as lower compared to the boys. Moreover, boys receive more attention and are more often verbally addressed. In line with this, McCullough (2002) described physics teachers as more often calling on boys than girls, approving of challenging remarks from boys but not from girls, and putting more demanding questions to boys than girls. Under certain circumstances, physics teachers' gender-specific evaluations and behavior can even manifest in significant differences in the grades they give to female vs. male students (Hofer, 2015). Such unfavorable conditions

correspond to the underachieving girls' low self-concept and interest in physics and may contribute to the development of underachievement.

Unbiased instruction and assessment that focus more on conceptual understanding than on memorizing and practicing formulae application, link to prior experiences, and support discussion can be expected to increase the underachieving girls' self-concept and interest in physics and positively influence performance (c.f. Häussler & Hoffmann, 2002; Hofer et al., 2015; Hoffmann, 2002; Hulleman & Harackiewicz, 2009; Lorenzo, Crouch, & Mazur, 2006; Siegle, Rubenstein, & Mitchell, 2014; Zohar, 2006; Zohar & Sela, 2003). Importantly, such kind of instruction and assessment that aim at conceptual understanding seem to benefit not only female underachievers but all learners (see Hofer et al., 2015).

Implications for Research and Limitations of the Present Study

In the above section, we tried to derive some ideas how physics instruction and underachieving girls interact and what kinds of classroom interventions consequently may counteract underachievement. Future research may pick up on these suggestions and combine instructional interventions with underachievement research in physics, as already done by Hofer and colleagues (2015).

As became clear, the statistical approach used to operationalize underachievement offers a useful framework for analyzing a broad variety of research questions related to underachievement in physics or any other domain. The potential-achievement discrepancy can be described probabilistically with latent profiles, allowing the uncertainty in categorizing the students to be explicitly factored in computationally. Based on the present study's methodology, latent transition analyses may be conducted to illuminate the conditions and genesis of physics underachievement over time. At the moment, only referring to a cross-sectional correlational design, we cannot say how underachievement in physics develops. To obtain a more comprehensive picture of physics underachievement, a broader range of student variables, including affective and personality variables or the students' gender stereotype endorsement, may be added to the analysis. A closer look at some of the other student profiles, the male physics overachievers, for instance, may also provide valuable new insights regarding predictors of subject-specific school success.

So far, the results can only be generalized to Swiss Gymnasium students. Without further testing in other countries, there is no guarantee that the findings will hold true in the context of a different country and a different educational system, too.

Sample 2 consisted of 14 classrooms. The number of classrooms and the class sizes, however, were too small to statistically consider the dependency in the data (see Muthén & Satorra, 1995; Wu & Kwok, 2012). At least 30 classrooms encompassing at least 30 students have been recommended to reliably perform such analyses (see Hox, 1998). Yet, the finding that the profile structure detected in sample 2 did not differ significantly from the profile structure detected in sample 1 suggests that classroom-level effects on the profiles may be neglected.

To conclude, the present study contributes to our understanding of the gender gap in physics. More than 50% of all boys belonged to the physics high achievers profile and the proportion of all girls belonging to this profile could be comparably high were it not for the group of females who have the intellectual potential to excel in physics but perform poorly in terms of grades. The statistical method applied in this study to investigate physics underachievement may help to gain further insights into the development and prevention of these girls' underachievement in physics.

References

- Adams, C. M. (1996). Gifted girls and science: Revisiting the issues. *Prufrock Journal*, 7(4), 447–458. <http://doi.org/10.1177/1932202X9600700404>
- Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 2(1), 1–14. <http://doi.org/10.1103/PhysRevSTPER.2.010101>
- Asparouhov, T., & Muthén, B. (2012). Auxiliary variables in mixture modeling: A 3-step approach using Mplus. *Mplus Web Notes*, 15. Retrieved from <http://statmodel2.com/examples/webnotes/webnote15.pdf>
- Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, 8(2), 125–145.
- Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Bucik, V., & Neubauer, A. C. (1996). Bimodality in the Berlin model of intelligence structure (BIS): A replication study. *Personality and Individual Differences*, 21(6), 987–1005. [http://doi.org/10.1016/S0191-8869\(96\)00129-8](http://doi.org/10.1016/S0191-8869(96)00129-8)
- Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3), 75–141. <http://doi.org/10.1177/1529100614541236>
- Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, 135(2), 218–261. <http://doi.org/10.1037/a0014412>
- Colangelo, N., Kerr, B., Christensen, P., & Maxey, J. (1993). A comparison of gifted underachievers and gifted high achievers. *Gifted Child Quarterly*, 37(4), 155–160. <http://doi.org/10.1177/001698629303700404>
- Conklin, A. M. (1940). *Failures of highly intelligent pupils. A study of their behavior by means of the control group*. New York: Bureau Of Publications, Teachers College, Columbia University.
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21. <http://doi.org/10.1016/j.intell.2006.02.001>

- Debacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *The Journal of Educational Research*, 93(4), 245–254. <http://doi.org/10.1080/00220670009598713>
- Dowdall, C. B., & Colangelo, N. (1982). Underachieving gifted students: Review and implications. *Gifted Child Quarterly*, 26(4), 179–184. <http://doi.org/10.1177/001698628202600406>
- Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. A primer. *Journal of Cross-Cultural Psychology*, 34(2), 195–210. <http://doi.org/10.1177/0022022102250427>
- Finch, W. H., & Bronk, K. C. (2011). Conducting confirmatory latent class analysis using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(1), 132–151. <http://doi.org/10.1080/10705511.2011.532732>
- Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., ... Pekrun, R. (2009). *PISA 2006 Skalenhandbuch. Dokumentation der Erhebungsinstrumente [PISA 2006 handbook of scales. Documentation of assessment instruments]*. Münster: Waxmann.
- Gagné, F. (2004). Transforming gifts into talents: The DMGT as a developmental theory. *High Ability Studies*, 15(2), 119–147. <http://doi.org/10.1080/1359813042000314682>
- Gagné, F. (2005). From gifts to talents. The DMGT as a developmental model. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 98–119). Cambridge, UK: Cambridge University Press.
- Geiser, C. (2011). *Data analysis with Mplus*. New York: Guilford Press.
- Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, 24(3), 229–252. <http://doi.org/10.1007/BF02289845>
- Gollwitzer, M. (2012). Latent-class-analysis. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 295–323). Berlin: Springer. Retrieved from http://link.springer.com/chapter/10.1007%2F978-3-642-20072-4_12
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, 100(1), 105–122. <http://doi.org/10.1037/0022-0663.100.1.105>
- Hart, C. (1996). Changing physics to suit the girls? In P. F. Murphy & C. V. Gipps (Eds.), *Equity in the classroom* (pp. 236–241). London and Washington, DC: Falmer Press and UNESCO.
- Häussler, P., & Hoffmann, L. (2002). An intervention study to enhance girls' interest, self-concept, and achievement in physics classes. *Journal of Research in Science Teaching*, 39(9), 870–888. <http://doi.org/10.1002/tea.10048>
- Heilbronner, N. N. (2012). The STEM pathway for women: What has changed? *Gifted Child Quarterly*, 57(1), 39–55. <http://doi.org/10.1177/0016986212460085>

- Hofer, S. I. (2015). *Studying gender bias in physics grading: The role of teaching experience and country*. Manuscript submitted for publication.
- Hofer, S. I., Stern, E., Rubin, H., & Schumacher, R. (2015). *Fostering conceptual understanding with cognitively activating instruction in physics classrooms: Evidence for general effects and special benefits for high potential students*. Manuscript in preparation.
- Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction*, 12(4), 447–465. [http://doi.org/10.1016/S0959-4752\(01\)00010-X](http://doi.org/10.1016/S0959-4752(01)00010-X)
- Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-72087-1_17
- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, 326(5958), 1410–1412.
- Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. BIS-Test, Form 4*. Göttingen: Hogrefe.
- 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. <http://doi.org/10.1016/j.lindif.2013.12.003>
- Kahle, J. B., & Lakes, M. K. (1983). The myth of equality in science classrooms. *Journal of Research in Science Teaching*, 20(2), 131–140. <http://doi.org/10.1002/tea.3660200205>
- Kang, N.-H., & Wallace, C. S. (2005). Secondary science teachers' use of laboratory activities: Linking epistemological beliefs, goals, and practices. *Science Education*, 89(1), 140–165. <http://doi.org/10.1002/sce.20013>
- 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(5), 448–470. <http://doi.org/10.2307/749801>
- Langer Tesfaye, C., & White, S. (2012). *High school physics teacher preparation*. American Institute of Physics Statistical Research Center.
- Lau, K.-L., & Chan, D. W. (2001). Identification of underachievers in Hong Kong: Do different methods select different underachievers? *Educational Studies*, 27(2), 187–200. <http://doi.org/10.1080/03055690120050419>
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2), 118–122. <http://doi.org/10.1119/1.2162549>
- Lubinski, D., & Benbow, C. P. (1992). Gender differences in abilities and preferences among the gifted: Implications for the math-science pipeline. *Current Directions in Psychological Science*, 1(2), 61–66. <http://doi.org/10.1111/1467-8721.ep11509746>

- Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, 1(4), 316–345. <http://doi.org/10.1111/j.1745-6916.2006.00019.x>
- Lubinski, D., & Benbow, C. P. (2007). Sex differences in personal attributes for the development of scientific expertise. In S. J. Ceci & W. M. Williams (Eds.), *Why aren't more women in science: Top researchers debate the evidence* (pp. 79–100). Washington, DC: American Psychological Association.
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280–295. <http://doi.org/10.1037/0022-0663.79.3.280>
- Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, 1(2), 133–163. <http://doi.org/10.1111/j.1745-6916.2006.00010.x>
- Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. S. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person- and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(2), 191–225. <http://doi.org/10.1080/10705510902751010>
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. <http://doi.org/10.1111/j.1467-8624.2005.00853.x>
- McCall, R. B. (1994). Academic underachievers. *Current Directions in Psychological Science*, 3(1), 15–19. <http://doi.org/10.1111/1467-8721.ep10769838>
- McCoach, D. B., & Siegle, D. (2003). The school attitude assessment survey-revised: A new instrument to identify academically able students who underachieve. *Educational and Psychological Measurement*, 63(3), 414–429. <http://doi.org/10.1177/0013164403063003005>
- McCullough, L. (2002). Women in physics: A review. *The Physics Teacher*, 40(2), 86–91. <http://doi.org/10.1119/1.1457312>
- Murphy, P., & Whitelegg, E. (2006a). Girls and physics: Continuing barriers to “belonging.” *Curriculum Journal*, 17(3), 281–305. <http://doi.org/10.1080/09585170600909753>
- Murphy, P., & Whitelegg, E. (2006b). Girls in the physics classroom: A review of the research on the participation of girls in physics. *Institute of Physics, London, UK*. Retrieved from http://oro.open.ac.uk/6499/1/Girls_and_Physics_Report.pdf
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology 1995, Vol 25*, 25, 267–316. <http://doi.org/10.2307/271070>
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (Seventh Edition). Los Angeles, CA: Muthén & Muthén. Retrieved from

- http://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4), 535–569. <http://doi.org/10.1080/10705510701575396>
- Organisation for Economic Co-operation and Development (2009). *Top of the class: High performers in science in PISA 2006*. OECD Publishing. Retrieved from <http://www.oecd-ilibrary.org/docserver/download/9809061e.pdf?expires=1394711955&id=id&accname=ocid72024074a&checksum=FBF7E1D665EB5EFDB3C197D19CE9D2D7>
- Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology*, 106(3), 696–710. <http://doi.org/10.1037/a0036006>
- Poorthuis, A. M. G., Juvonen, J., Thomaes, S., Denissen, Jaap J. A., Orobio de Castro, Bram, & van Aken, Marcel A. G. (2014). Do grades shape students' school engagement? The psychological consequences of report card grades at the beginning of secondary school. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000002>
- Raven, J. C., Raven, J., & Court, J. H. (1992). *Raven's Progressive Matrices und Vocabulary Scales. Teil 4 Advanced Progressive Matrices*. (S. Bulheller & H. Häcker, Trans.). Frankfurt: Swets & Zeitlinger.
- Reis, S. M. (1991). The need for clarification in research designed to examine gender differences in achievement and accomplishment. *Roeper Review*, 13(4), 193–198. <http://doi.org/10.1080/02783199109553357>
- Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted Child Quarterly*, 44(3), 152–170. <http://doi.org/10.1177/001698620004400302>
- Robertson, K. F., Smeets, S., Lubinski, D., & Benbow, C. P. (2010). Beyond the threshold hypothesis: Even among the gifted and top math/science graduate students, cognitive abilities, vocational interests, and lifestyle preferences matter for career choice, performance, and persistence. *Current Directions in Psychological Science*, 19(6), 346–351. <http://doi.org/10.1177/0963721410391442>
- Rost, D. H., Sparfeldt, J. R., & Schilling, S. R. (2007). *Differentielles schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten (DISK-Gitter mit SKSLF-8)*. Manual. Göttingen: Hogrefe.
- Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest*

- in learning and development* (pp. 183–212). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Schober, B., Reimann, R., & Wagner, P. (2004). Is research on gender-specific underachievement in gifted girls an obsolete topic? New findings on an often discussed issue. *High Ability Studies*, 15(1), 43–62. <http://doi.org/10.1080/1359813042000225339>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333–343.
- Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learning Environments Research*, 9(3), 253–271. <http://doi.org/10.1007/s10984-006-9012-x>
- Seidel, T., Prenzel, M., Rimmel, R., Dalehefte, I. M., Herweg, C., Kobarg, M., & Schwindt, K. (2006). Blicke auf den Physikunterricht. Ergebnisse der IPN Videostudie. *Zeitschrift für Pädagogik*, 52(6), 799–821.
- Siegle, D. (2013). *The underachieving gifted child: Recognizing, understanding, and reversing underachievement*. Waco, TX: Prufrock Press.
- Siegle, D., Rubenstein, L. D., & Mitchell, M. S. (2014). Honors students’ perceptions of their high school experiences the influence of teachers on student motivation. *Gifted Child Quarterly*, 58(1), 35–50. <http://doi.org/10.1177/0016986213513496>
- Smith, E. (2003). Failing boys and moral panics: Perspectives on the underachievement debate. *British Journal of Educational Studies*, 51(3), 282–295. <http://doi.org/10.1111/1467-8527.t01-2-00239>
- Snyder, K. E., & Linnenbrink-Garcia, L. (2013). A developmental, person-centered approach to exploring multiple motivational pathways in gifted underachievement. *Educational Psychologist*, 48(4), 209–228. <http://doi.org/10.1080/00461520.2013.835597>
- Sparfeldt, J. R., Schilling, S. R., & Rost, D. H. (2006). Hochbegabte Underachiever als Jugendliche und junge Erwachsene. *Zeitschrift Für Pädagogische Psychologie*, 20(3), 213–224. <http://doi.org/10.1024/1010-0652.20.3.213>
- Specht, J., Luhmann, M., & Geiser, C. (2014). On the consistency of personality types across adulthood: Latent profile analyses in two large-scale panel studies. *Journal of Personality and Social Psychology*, 107(3), 540–556.
- Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, 135(6), 859–884. <http://doi.org/10.1037/a0017364>
- Taasoobshirazi, G., & Carr, M. (2008). Gender differences in science: An expertise perspective. *Educational Psychology Review*, 20(2), 149–169. <http://doi.org/10.1007/s10648-007-9067-y>

- Taconis, R. (1995). *Understanding based problem solving: Towards qualification-oriented teaching and learning in physics education*. Technische Universiteit Eindhoven.
- Taconis, R., Ferguson-Hessler, M. G. M., & Broekkamp, H. (2001). Teaching science problem solving: An overview of experimental work. *Journal of Research in Science Teaching*, 38(4), 442–468. <http://doi.org/10.1002/tea.1013>
- Thorndike, R. L. (1963). *The concepts of over and underachievement*. New York: Bureau Of Publications, Teachers College, Columbia University.
- Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, North Carolina: Information Age Publishing.
- UCLA: Statistical Consulting Group (2014, July 18). Mplus FAQ. How can I compute a chi-square test for nested models with the MLR or MLM estimators? Retrieved July 18, 2014, from http://www.ats.ucla.edu/stat/mplus/faq/s_b_chi2.htm
- Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars, & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press.
- Von Maurice, J., Dörfler, T., & Artelt, C. (2014). The relation between interests and grades: Path analyses in primary school age. *International Journal of Educational Research*, 64, 1–11. <http://doi.org/10.1016/j.ijer.2013.09.011>
- Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, 140(4), 1174–1204. <http://doi.org/10.1037/a0036620>
- Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4), 817–835. <http://doi.org/10.1037/a0016127>
- Wu, J.-Y., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(1), 16–35. <http://doi.org/10.1080/10705511.2012.634703>
- Yang, C.-C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50(4), 1090–1104. <http://doi.org/10.1016/j.csda.2004.11.004>
- Yang, C.-C., & Yang, C.-C. (2007). Separating latent classes by information criteria. *Journal of Classification*, 24(2), 183–203. <http://doi.org/10.1007/s00357-007-0010-1>
- Ziegler, A., Ziegler, A., & Stoeger, H. (2012). Shortcomings of the IQ-based construct of underachievement. *Roepers Review*, 34(2), 123–132. <http://doi.org/10.1080/02783193.2012.660726>
- Zohar, A. (2006). Connected knowledge in science and mathematics education. *International Journal of Science Education*, 28(13), 1579–1599. <http://doi.org/10.1080/09500690500439199>

Zohar, A., & Sela, D. (2003). Her physics, his physics: Gender issues in Israeli advanced placement physics classes. *International Journal of Science Education*, 25(2), 245–268. <http://doi.org/10.1080/09500690210126766>

Acknowledgements

We wish to thank Jessica Büetiger for her great assistance throughout the project and Nora Müller, Lara Ringier, Pauline Hofer, and Sebastian Seehars for their support in coding huge amounts of test data. We would also like to thank Bruno Rütsche and Peter Edelsbrunner for their help with technical and software issues.

3. Studying Gender Bias in Physics Grading: The Role of Teaching Experience and Country

Sarah I. Hofer

The existence of gender-STEM (science, technology, engineering, and mathematics) stereotypes has been repeatedly documented. This article examines physics teachers' gender bias in grading and the influence of teaching experience in Switzerland, Austria, and Germany. In a 2×2 between-subjects design, with years of teaching experience included as moderating variable, physics teachers ($N = 780$) from Switzerland, Austria, and Germany graded a fictive student's answer to a physics test question. While the answer was exactly the same for each teacher, only the student's gender and specialization in languages vs. science were manipulated. Specialization was included to gauge the relative strength of potential gender bias effects. Multiple group regression analyses, with the grade that was awarded as the dependent variable, revealed only partial cross-border generalizability of the effect pattern. While the overall results in fact indicated the existence of a consistent and clear gender bias against girls in the first part of physics teachers' careers that disappears with increasing teaching experience for Swiss teachers, Austrian teachers, and German female teachers, German male teachers showed no gender bias effects at all. The results are discussed regarding their relevance for educational practice and research.

Keywords: *Gender bias; Teaching experience; Physics instruction*

Introduction

Secondary school has been identified as crucial point in time to consolidate the gender gap in engagement, interest, and participation in STEM (science, technology, engineering, and mathematics) fields (Ceci, Williams, & Barnett, 2009). Among all of the explanations that are provided for the gender gap, the present paper addresses the basic aspect of gender biased grading in physics. A number of studies examined accuracy and various biases in teachers' judgments of student performance (e.g., Dünnebier, Gräsel, & Krolak-Schwerdt, 2009; Glock & Krolak-Schwerdt, 2014; Südkamp, Kaiser, & Möller, 2012). There is no recent work, however, that explicitly investigates whether secondary school teachers' grading in physics indeed reveals a bias to the detriment of girls. The present study hence aims to fill this gap and additionally shed light on the role of teaching experience. The generalizability of potential gender bias effects in secondary school physics is examined by comparing teachers' bias patterns across three German-speaking countries that are culturally closely related.

In the following sections, the literature addressing gender bias in teachers' judgments in STEM fields is outlined. First, mechanisms that underlie biased judgments and the potential influence of teaching experience are considered. Then I turn to existing research on gender bias in academic judgments. Finally, the cross-border generalizability of gender bias effects is briefly addressed, before the present study is introduced.

Gender Bias in Teachers' Judgments in STEM Fields

There are only a few studies that focus on gender bias in teachers' judgments in the specific domain of physics. Therefore, most of the findings and theoretical considerations that are summarized in the following sections relate to the broader category of STEM fields.

Underlying Mechanisms: Gender-STEM Stereotypes

To be able to navigate through our highly demanding social environment, schemata are applied that efficiently categorize our perceptions (Bartlett, 1932). Schemata that refer to members of social groups are stereotypes. A stereotype associates a social group concept with one or a set of attribute concepts (e.g., Greenwald et al., 2002). The most acknowledged

models that have been proposed to explain the influence of stereotypes on judgment processes are dual process models or continuum models which are characterized by additional intermediate processes (see Brewer, 1988; S. T. Fiske & Neuberg, 1990) and parallel-constraint-satisfaction models (see Kunda & Thagard, 1996). All models, however, arrive at very similar conclusions in terms of factors that are expected to affect the extent of a stereotype's influence on the judgment process. Accordingly, among others, cognitive business or limited cognitive resources in the judgment situation and ambiguous information can increase the probability that stereotypes take effect and dominate individuating information (e.g., Chaiken & Maheswaran, 1994; Kunda & Spencer, 2003; Kunda & Thagard, 1996).

In the present study, gender-STEM stereotypes are expected to potentially bias teachers' judgments. In general, gender stereotypes to some extent reflect but also contribute to existing gender differences in behavior (see Eagly & Wood, 2013). Perceived incongruity between gender stereotypes and stereotypic job roles may lead to biased evaluations and prejudice against those females (or males) performing in a nontraditional domain (e.g., Eagly & Koenig, 2008; Eagly, Wood, & Diekmann, 2000). In line with this, existing research points to a commonly perceived mismatch between stereotypic views of women, on the one hand, and scientists, on the other hand (see Farenga & Joyce, 1999; Kessels, Rau, & Hannover, 2006; Nosek et al., 2009; Nosek, Banaji, & Greenwald, 2002). Accordingly, gender-STEM stereotypes can be defined as stronger associations between STEM-related content and males than females (see Miller, Eagly, & Linn, 2014; Nosek et al., 2007, 2002). There is good reason to assume that such more general gender-STEM stereotypes also apply to the more specific domain of physics that is part of the STEM fields. In the gender-science Implicit Association Test that is aimed to measure the strength of the stereotypic association of science with men, physics is presented as one instantiation of science words (e.g., Nosek et al., 2009). When comparing physics and math teachers' implicit theories about their students' achievement and ability in their respective fields, physics teachers' cognition tended to be even slightly more gender-biased in favor of boys (Heller, Finsterwald, & Ziegler, 2010).

To sum up, when physics teachers evaluate the performance of students, gender-STEM stereotypes may influence the judgment process, especially in judgment situations that are cognitively demanding and provide ambiguous information.

The Role of Teaching Experience

In the classroom, the information that is available to make a decision about a student's performance level usually is ambiguous and open to various interpretations. The accuracy of teachers' ratings of students' performance indeed seems to be lower for science and social studies than for reading, language arts, or mathematics (Hopkins, George, & Williams, 1985) and lower for conceptual questions than for computational questions (Coladarci, 1986), which inherently provide less strict evaluation criteria and more interpretative ambiguity.

Both the perceived ambiguity of information and a high demand for cognitive resources in the judgment situation can be expected to diminish with increasing teaching experience. There is evidence that expert teachers, in comparison to novices, are able to automatize parts of their work (see Carter, Sabers, Cushing, Pinnegar, & Berliner, 1987; Leinhardt & Greeno, 1986), and to quickly and correctly recognize more meaningful patterns as a function of their experience (see Berliner, 2001). Expert teachers, but not novices, seem to use elaborated schemata as frameworks to efficiently interpret and understand the often complex information that has to be processed (Carter et al., 1987). Although, in general, mere experience is not sufficient to determine expert teachers (see Palmer, Stough, Burdenski, Jr., & Gonzales, 2005), these skills are suggested to develop with increasing teaching experience.

Accordingly, the need to invoke stereotypes in grading may also decrease with increasing teaching experience, what is supported by the following findings. Krolak-Schwerdt, Böhmer, and Gräsel (2009, 2012) instructed participants to read students' case reports and to either form an impression of the students' behavior and performance or to predict future performance. The latter was stressed to be relevant for the student's future academic career. The authors found that teachers with at least ten years of teaching experience but not laymen (students of natural sciences) were able to flexibly switch from a category-based processing to a processing of relevant individuating information when they had to predict students' performance. In a related study (Dünnebier et al., 2009), student performance judgments of student teachers were more influenced by prior information than judgments of teachers with at least eight years of experience. Finally, Babad (1985) found that elementary school teachers' grading in the context of text comprehension varied significantly as a function of the fictitious performance label (excellent vs. weak student) in the group of the less experienced teachers (not more than eight years of experience), but not in the group of the more experienced teachers.

Existing Research: Disentangling Bias and Accuracy

There are two main approaches that have dominated research on teachers' judgment biases. In the first approach, the characteristics that are expected to trigger biased evaluations in a particular judgment domain are manipulated, while the content that has to be judged stays the same in each condition. Focusing on potential gender-STEM bias effects, Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman (2012) applied this approach to the educational domain and found that science faculty staff derived significantly higher competence levels from identical application materials with a male name than those with a female name. By investigating secondary school science teachers' evaluations of the same essays that were either indicated to originate from a girl or a boy, Goddard Spear (1984a) also reported a rather consistent bias towards boys, with regard to grades, estimated competence, and the students' perceived inclination for science (see also Goddard Spear, 1984b). Although the author used a similar design, Baird (1998) did not find any gender bias in grading for A-level examinations in chemistry.

In the second, correlational approach, teachers' judgments of student performance are related to objective performance measures to estimate judgment accuracy and biases. Judgments that are influenced by stereotypes are regarded as accurate or biased depending on the degree they reflect actual group differences (see e.g., Jussim & Eccles, 1992; Madon et al., 1998). There is evidence that teachers tend to overestimate their male students' proficiency in math when actual performance is accounted for (Jussim & Eccles, 1992; Robinson-Cimpian, Lubinski, Ganley, & Copur-Gencturk, 2014). In keeping with Robinson-Cimpian and colleagues (2014), equally performing girls have to outmatch boys in terms of teachers' perceived effort, diligence, and manners to be rated as equally proficient in math.

To sum up, there is some evidence for a bias against females in STEM fields. Most of the research up to now, however, addressed science in general or math, but not physics. Moreover, to rule out the influence of small but existent self-fulfilling prophecy effects (see Jussim & Harber, 2005) and stereotype threat effects (e.g., Nguyen & Ryan, 2008) on student performance measures, the highly controlled experimental approach may be preferred to the correlational approach when no conclusions about teachers' judgment accuracy are intended. Accordingly, in the present study that applied the experimental approach, 'bias' does not refer to a systematic deviation from objective performance assessments but is simply meant

to indicate a systematic variation in teachers' judgments as a function of the experimental variation of a stereotyped characteristic.

Cross-Border Generalizability of Gender-STEM Bias Effects

Overall, more than 70% of the participants in a study by Nosek and colleagues (2009) from 34 countries all over the world hold implicit gender-STEM stereotypes. The degree of national stereotype endorsement turned out to predict nation-level gender achievement gaps in school science (Nosek et al., 2009). Also the proportion of women who participate in tertiary science education predicts the degree of nation-level gender-STEM stereotype endorsement (Miller et al., 2014). On a general level, the cultural context shapes the categories that are used to organize our perceptions and hence also influences the content of stereotypes (see e.g., A. P. Fiske, Kitayama, Markus, & Nisbett, 1998). Gender-STEM bias effects in teachers' judgments may thus generalize over countries that are culturally closely related and that are comparable in terms of the nation-level representation of women in STEM fields and in terms of gender differences in science performance measures.

The Present Study

The present study applied the experimental approach to examine gender bias in physics teachers' judgments and the role of teaching experience. Secondary school physics teachers received a physics test question and the same written student answer, accompanied by the prompt to assign a grade. The physics test question asked a fictive student for a written explanation about his or her conceptual understanding of Newtonian mechanics. Compared to problems that require computation, conceptual items that are expected to imply higher ambiguity and leeway in construal were considered to be a more interesting evaluation situation to examine. Two factors were manipulated in a short introductory text: student gender and specialization in languages vs. science. The second factor, specialization, was only included to gauge the relative strength of potential gender bias effects. Effects of gender on grading could then be compared to the effects of another category (students focusing on languages vs. students focusing on science) that is assumed to more clearly reflect actual group differences but represents a less prevailing and less distinct social category.

Based on existing research, the present study expected physics teachers to show a gender bias in grading, to the detriment of girls. Student gender was assumed to more strongly influence grading than the less prominent social category student specialization. The study further aimed to investigate the potential moderating effect of teaching experience, which may reduce gender bias with increasing years of practice. In comparison to most other studies that contrasted groups of less and more experienced teachers, this study included teaching experience as continuous variable. Because the three German-speaking countries Switzerland, Austria, and Germany are culturally closely related and comparable in terms of the nation-level representation of women in STEM fields (e.g., European Commission, 2013) and in terms of an existing advantage for boys in science performance measures (e.g., Organisation for Economic Co-operation and Development, 2011), a generally valid pattern of bias effects independent of German-speaking country was expected.

Method

Design

This study applied a 2×2 between-subjects factorial design. The two independent variables were student gender (female vs. male) and student specialization (languages vs. science). The grade that teachers assigned to the fictive student test answer was the dependent variable. The dependent variable was measured with only one item because in real grading situations, a student's oral or written answer is generally evaluated by assigning a single grade. Hence, asking the teachers to assign a grade ascertained ecological validity and allowed fast and intuitive processing of the survey. Teaching experience in years served as continuous moderating variable to investigate the influence of teaching experience on the effects of gender and specialization on the grade that was awarded. The effect pattern was compared between samples from Switzerland, Austria, and Germany to be able to examine its generalizability.

The study was run through the use of an online-survey tool, SoSciSurvey (Leiner, 2014), which could be accessed from every web-enabled device via a link. Physics teachers' associations and science education research institutions in Switzerland, Austria, and Germany were contacted and asked to distribute a request for participation that included the survey link

to their mailing lists. The mailing lists explicitly addressed physics teachers. In the request for participation in the email and in the survey itself, it was emphasized that the study was exclusively aimed at physics teachers. Three country-specific links and surveys were prepared. Certain demographic and personal questions, as well as the grading system, were adapted to the countries' respective national standards. Both in the request for participation and in the introductory text in the survey itself, the overall objective of the study was described as investigating the process of performance evaluation in secondary school physics. The research interest in gender bias, however, was not made explicit in order to reduce social desirability biases and conscious efforts to avoid prejudice that could have, otherwise, distorted the findings. Hence, teachers were told that this research project particularly aimed to examine the correspondence between two assessment situations: A student's performance on a test was assessed either by each test question being evaluated by a different physics teaching expert or by one expert evaluating the complete test. This cover story also justified why they were asked to evaluate a single test answer by assigning a grade. Because the teachers were told that they were evaluating a real student's test answers that were provided by different schools, this study examined gender bias in an experimental design, while maintaining good ecological validity.

Teacher Samples

A sample size of 20 physics teachers per experimental condition, which resulted in at least 80 teachers per country, was set as the lower limit. Country-specific data collection was finished after this limit was reached and when the survey was not accessed for at least four days. Following this procedure, 167 cases were initially registered from Switzerland, 178 from Austria, and 589 from Germany. In all of the three German-speaking countries, physics is more intensively instructed only in the higher tracks of secondary school. To arrive at comparable samples, only those participants who indicated that they taught at a higher level secondary school were considered. Participants whose age suggested that they had already retired were further excluded from the analyses. When no grade was awarded, the participant's data were eliminated. By checking IP-addresses and personal data, multiple completions of the survey were detected, and the respective data were deleted. Hence, the Swiss sample finally included $N = 116$ (14 women) physics teachers. On average, they were $M = 48.83$ ($SD = 9.26$) years old and had $M = 18.32$ ($SD = 10.20$) years of teaching

experience. Due to the multilingualism in Switzerland, German language proficiency was additionally collected at the beginning of the survey in order to directly exclude teachers who did not have a German-speaking background. The Austrian sample included $N = 137$ (59 women) teachers, with a mean age of $M = 47.03$ ($SD = 10.89$) years and a mean length of teaching experience of $M = 19.58$ ($SD = 12.40$) years. The German sample included $N = 527$ (125 women) physics teachers, with a mean age of $M = 46.64$ ($SD = 10.96$) years and a mean length of teaching experience of $M = 17.17$ ($SD = 11.84$) years. The gender distribution in the three samples closely resembled country-specific statistical information on the gender distribution of physics teachers. The total sample included $N = 780$ German-speaking secondary school physics teachers.

Procedure

When accessing the online-survey, a brief introductory text informed the participants about the study's aim and the procedure. After the anonymous assessment of demographic and personal information, including years of teaching experience, participants were randomly forwarded to one of the four conditions. In all of the four conditions, teachers received exactly the same information, with the exception of all of the terms that referred to a fictive student's gender and the student's specialization (languages vs. science), which were interchanged based on the condition. Following a short text that introduced the student, the teachers saw the physics test question, which asked the fictive student for a written explanation that targeted his or her conceptual understanding of Newton's axioms, and the answer of the student. The teachers were asked to evaluate the student's answer by assigning a grade according to their respective, country-specific school grading systems. Answers were graded by moving a continuous slider that instantaneously provided the corresponding number of the grade to one decimal point. Due to the randomization of the experimental conditions, systematic individual differences in performance judgment severity or leniency could be neglected. For illustrative purposes, essential parts of the German online-survey were translated into English and are summarized in Figure 3.1.

The test questions, student answers, and brief descriptions of the context are directly provided by the participating schools. We only summarize the information that we receive. The content that is assessed in the test questions had always been taught in the lessons before.

In the following text, you will see a test question on Newtonian mechanics and a _____ (female/male student's) answer. _____ (She/He) is in _____ (her/his) Junior Year and takes Honors/AP courses. During _____ (her/his) school career, _____ (she/he) has focused on _____ (languages/the natural sciences), thus far. Please evaluate the _____ (female/male student's) answer.

_____ (She/He) was asked the following test question:

Two skateboarders who significantly differ in their masses each stand on a skateboard, face to face. They are connected by a tensioned rope. The left and lighter skateboarder actively pulls the rope, while the heavier right skateboarder only holds it. What happens? Explain your assumption in approximately five to six sentences. Friction is negligible.

The _____ (female/male student's) answer:

In general, force is composed of a person's mass and acceleration. The right skateboarder has to hold the rope as strongly as the left skateboarder pulls it. Both of the skateboarders are, thus, affected by forces of equal strengths, although only the left skateboarder pulls the rope. Consequently, nothing should happen because the two forces cancel out one another. Because the mass of the left skateboarder, however, is smaller than the mass of the right skateboarder, the left skateboarder has a higher acceleration than the right skateboarder. As a result, the left skateboarder most likely should at least move a small amount in the direction of the right skateboarder.

Please evaluate this _____ (female/male student's) answer by assigning a grade on a scale from A, with a corresponding grade point of 4.0, to F, with a corresponding grade point of 0.0.

In order to do this, please move the slider to the desired position.

Figure 3.1. English adaptation of the instructions and information that the teachers received. Terms that were interchanged in the four conditions are omitted, and variants are presented in parentheses. Note that in the German language, the student's gender is simply indicated by slightly changing the word's ending (female student = Schülerin, male student = Schüler).

The Judgment Situation

In this study, a student's answer to one conceptual question was used as the judgment situation. This judgment situation was considered to be a good proxy, both for the evaluation of a whole test and the evaluation of a student's classroom participation over a certain period

of time because the same stereotypic beliefs may also influence the teacher's interpretation of a student's classroom contributions. The conceptual test question that was used in this study was adapted from the basic Mechanics Concept Test (bMCT), a Rasch-scaled multiple-choice test on the conceptual understanding of Newton's three axioms. The "skateboarder question" (see Figure 3.1) was chosen because it covers a problem that, in fact, frequently appears in physics textbooks, exams, and classroom instruction in all of the three countries and requires a complex answer that potentially includes several correct and incorrect statements. In the process of the bMCT's development, a variety of oral and written student answers to this conceptual question were recorded and used to design three exemplary student answers. The aim was to arrive at an answer that represented average student performance and was neither completely wrong nor absolutely correct, in order to leave room for interpretation. The three answers were given to five informed physics teaching experts, who were asked to assign a grade to each of them. The answer that most unequivocally reflected average performance was finally chosen and used in the study (see Figure 3.1).

Data Analysis

To investigate a potential gender bias in physics grading and the influence of teaching experience within and across the three countries, multiple group regression analyses were performed with Mplus Version 7.11 (Muthén & Muthén, 2012) with country as grouping variable. Grades were transformed into z-scores for each country in order to account for the different grading scales and, if necessary, recoded to create a grade scale where higher values indicated higher performance. This joint grade scale is referred to in the following section and used in the analyses.

Grades were regressed on student gender (0 = female, 1 = male) and specialization (0 = languages, 1 = science). Teaching experience, the interaction between gender and teaching experience, as well as the interaction between specialization and teaching experience, were further included as predictors to be able to examine the potential moderating effect of teaching experience. Teaching experience, which was measured in years, was entered into the regression without being z-standardized to be able to examine the potential change in the gender bias effect with growing years of teaching experience. Consequently, the regression coefficient of gender reflected the influence of a fictive student's gender on the grades that were awarded at the beginning of the teaching career – with zero teaching experience.

To gain further insights into the meaning of potential interaction effects between the fictive student's gender and teaching experience in the empirical, and not linearly modelled, data, an additional analysis was performed. Grades were averaged within bins of five years of teaching experience, resulting in nine bins. Within each teaching experience bin, the mean grades that were awarded to a fictive female student were compared to the mean grades that were awarded to a fictive male student using t-tests.

Existing research suggests that the teacher's gender should have no influence on (gender) bias effects (see Moss-Racusin et al., 2012; Swim, Borgida, Maruyama, & Myers, 1989). Nevertheless, to rule out such influences, measurement invariance in terms of the regression model was investigated across the teachers' gender within each country separately. Only after the analysis of measurement invariance across the teachers' gender, which indicated whether female and male teachers from the same country could be reasonably considered together or had to be considered separately, the cross-border generalizability of the effect pattern was investigated.

One aim of this study was to examine the generalizability of potential gender bias effects, hypothesizing a generally valid pattern of effects. Consequently, the alternative hypothesis assumed differences in the patterns between countries or female and male physics teachers, respectively. When examining the cross-border generalizability of the effect pattern and measurement invariance in terms of the teachers' gender, a type II error in significance testing (assuming that there is no effect of country or the teachers' gender when there is an effect) may thus be regarded as more problematic than a type I error (assuming that there is an effect of country or the teachers' gender when there is none). Therefore, the significance level for all tests of invariance was set to $\alpha = .20$ to increase the test's power.

Results

In the original regression model, grades were regressed on student gender and specialization, teaching experience, the interaction between gender and teaching experience, as well as the interaction between specialization and teaching experience. The specialization of the fictive student as well as the interaction between specialization and teaching experience, however, turned out to have no systematic influence on the grade that was awarded (all $ps \geq .12$), neither for female nor for male teachers in none of the countries.

Therefore, specialization and the interaction between specialization and teaching experience were excluded from all analyses that are reported in the following sections.

Descriptive Statistics

Descriptive statistics that are related to the grade scale, without considering teaching experience, can be found in Table 3.1. Grade data is presented for each country separately organized according to the two experimentally manipulated variables.

Table 3.1

Country- and Condition-Specific Descriptive Statistics for the Grade Scale and Unstandardized Grades

Specialization	Gender					
	Female			Male		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
CH						
Languages	28	0.07 (4.07)	1.14 (0.92)	32	0.02 (4.03)	1.04 (0.84)
Science	27	-0.16 (3.89)	0.95 (0.77)	29	0.05 (4.05)	0.89 (0.72)
AU						
Languages	33	-0.04 (3.11)	1.09 (1.11)	32	0.11 (2.95)	1.03 (1.05)
Science	35	-0.20 (3.26)	0.92 (0.93)	37	0.13 (2.93)	0.97 (0.99)
GE						
Languages	126	-0.03 (3.32)	1.02 (1.07)	143	0.06 (3.22)	1.00 (1.05)
Science	125	0.06 (3.22)	0.96 (1.01)	133	-0.10 (3.39)	1.02 (1.06)

Note. The grade scale resulted from z-standardization within each country and recoding so that higher values indicated higher performance. Statistics for the unstandardized grades are in parentheses. In Switzerland (CH), grades range from 6 (best) to 1 (worst); in Austria (AU), grades range from 1 (best) to 5 (worst); and in Germany (GE), grades range from 1 (best) to 6 (worst).

Effects of the Teachers' Gender

The regression model, which now only included the three predictor variables gender, teaching experience, and the interaction between gender and teaching experience, was tested

in terms of measurement invariance across the teachers' gender within each country separately (i.e., multiple group regression analyses with the teachers' gender as grouping variable). In the unrestrictive model, the regression was estimated independently for female and male physics teachers within each country. In the restrictive model, the regression coefficients were constrained to be equal between female and male physics teachers of the same nationality assuming measurement invariance. The two nested models were compared using the sample-size adjusted Bayesian Information Criterion (aBIC; Sclove, 1987) and the standard Bayesian Information Criterion (BIC; Schwarz, 1978), where lower values indicate better model-fit, as well as log-likelihood tests. Using log-likelihood tests, more restrictive models are compared to less restrictive but nested models (for detailed information on the test, see UCLA: Statistical Consulting Group, 2014). In addition to log-likelihoods, scaling correction factors and the number of free parameters for the models that are compared have to be considered in order to calculate a chi-square distributed test statistic that yields a p -value ($LL\ p$). In the case of no significant discrepancies in model-fit, the more restrictive model suggesting measurement invariance should be chosen.

As regards Swiss female and male teachers, both the aBIC and BIC (restrictive: 333 and 355 vs. unrestrictive: 334 and 365) and the log-likelihood test ($LL\ p = .35$) indicated measurement invariance allowing a joint consideration. Yet, based on a sample of only 14 Swiss female teachers, gender differences in the effect pattern cannot be ruled out definitely until further research confirms this finding. Also in the Austrian sample, the aBIC and BIC (restrictive: 392 and 414 vs. unrestrictive: 394 and 426) as well as the log-likelihood test ($LL\ p = .40$) revealed measurement invariance across the teachers' gender. In the German sample, however, the results suggested differences in the effect patterns of female and male physics teachers. Although the aBIC and BIC (restrictive: 1502 and 1524 vs. unrestrictive: 1506 and 1538) again favored the restrictive model, the log-likelihood test indicated a better fit of the unrestrictive model ($LL\ p < .20$). Prompted by the outcome of the log-likelihood test, the German sample was split to be able to take into account even small differences in the effect patterns of female and male teachers.

Effects of Student Gender and Teaching Experience across Countries

Based on the analysis of effects of the teachers' gender, the German sample was divided into female and male physics teachers while in Switzerland and Austria female and male

teachers were considered together. Now effect patterns could be compared across countries and the two German subsamples. Three models were constructed. The most restrictive model, Model 1, suggested similar effects across all of the three countries. This model hence represented the original hypothesis that the bias effect pattern generalizes over all of the three countries and is generally valid. Model 3, by contrast, constituted the unrestrictive model that allowed for unique effect patterns within each country and the two German subsamples. If Model 3 proved to fit the data best, the effect patterns could be considered highly context-specific. An inspection of the Model 3 regression coefficients that were estimated independently within each country and the two German subsamples (see Table 3.2) suggested similar effects for Swiss teachers, Austrian teachers, and German female teachers but not for German male teachers. To be able to investigate the apparently divergent effect pattern of German male physics teachers, an additional model, Model 2, was constructed that consequently suggested similar effects across all of the three countries except for German male teachers. Model 2 thus represented a less strict version of the expected cross-border generalizability of gender-STEM bias effects (i.e., partial generalizability). No further models were constructed because these models sufficed to examine the generalizability of the effect pattern. Accordingly, regression analyses were run with the regression parameters constrained to be equal across all of the three countries (Model 1), across all of the three countries with the exception of the German male physics teachers that were freed (Model 2), and with all of the parameters estimated freely within each country and the two German subsamples (Model 3).

After the three models were estimated, their fit to the data was contrasted to find the model which best described the effect patterns across the countries including the two German subsamples. Hence, Model 1, Model 2, and Model 3 were compared, again using the aBIC and BIC as well as log-likelihood tests. The most restrictive Model 1 was accordingly compared to the less restrictive Model 2 as an alternative model. In a second step, Model 2 was compared to the unrestrictive Model 3 as an alternative model. In the case of no significant discrepancies in model-fit, the more restrictive, more parsimonious model should be chosen. In the case of significant differences, however, the more restrictive model fits the data significantly worse than the less restrictive model indicating that the less restrictive model should be chosen. The results of the multiple group regression analyses and the model comparisons are summarized in Table 3.2.

Table 3.2

Comparison of Three Multiple Group Regression Models that Predicted Grades Based on Gender, Teaching Experience, and the Interaction between Gender and Teaching Experience (gender \times exp)

Variable in country	Model 1 (CH=AU=GE)					Model 2 (CH=AU=GE females)					Model 3 (unrestrictive)				
	<i>b</i>	<i>SE</i>	<i>p</i>	ICs	<i>LL p</i>	<i>b</i>	<i>SE</i>	<i>p</i>	ICs	<i>LL p</i>	<i>b</i>	<i>SE</i>	<i>p</i>	ICs	<i>LL p</i>
				2233/ 2268	-				2229/ 2273	.00				2248/ 2311	.93
CH															
gender ^a	0.28	0.13	.03			0.77	0.18	.00			0.84	0.37	.02		
experience	0.00	0.00	.79			0.02	0.01	.03			0.02	0.02	.23		
gender×exp	-0.02	0.01	.02			-0.03	0.01	.00			-0.04	0.02	.04		
AU															
gender	0.28	0.13	.03			0.77	0.18	.00			0.86	0.28	.00		
experience	0.00	0.00	.79			0.02	0.01	.03			0.02	0.01	.04		
gender×exp	-0.02	0.01	.02			-0.03	0.01	.00			-0.03	0.01	.01		
GE females															
gender	0.28	0.13	.03			0.77	0.18	.00			0.64	0.32	.05		
experience	0.00	0.00	.79			0.02	0.01	.03			0.01	0.01	.54		
gender×exp	-0.02	0.01	.02			-0.03	0.01	.00			-0.03	0.02	.09		
GE males															
gender	0.28	0.13	.03			-0.10	0.18	.59			-0.10	0.18	.59		
experience	0.00	0.00	.79			-0.01	0.01	.03			-0.01	0.01	.03		
gender×exp	-0.02	0.01	.02			0.00	0.01	.78			0.00	0.01	.78		

Note. CH = Switzerland, AU = Austria, GE = Germany; *LL p* = *p*-values that resulted from the log-likelihood tests.

^a 0 = female, 1 = male.

In regard to the log-likelihood tests, Model 2 fitted the data significantly better than the most restrictive Model 1 ($LL\ p < .01$). The least restrictive Model 3 did not fit the data significantly better than the more restrictive Model 2 ($LL\ p = .93$). The aBIC further indicated the superiority of Model 2. Although the BIC favored the most restrictive Model 1, the BIC of Model 2 only slightly exceeded the value calculated for Model 1. Hence, Model 2, which suggested similar effect patterns across all of the three countries with the exception of the German male physics teachers, turned out to best describe the effect patterns across countries and the two German subsamples and is interpreted in the following.

According to Model 2 (see Table 3.2), the analysis revealed both a significant main effect of gender ($b_{\text{gender}} = 0.77$) and a clear moderating effect of teaching experience on the relationship between gender and grades ($b_{\text{gender} \times \text{exp}} = -0.03$) in the samples of Swiss, Austrian, and German female teachers. The gender effect that was reflected in an advantage of approximately 0.77 standard deviations on the grade scale for the fictive boy thus represented teachers' gender bias at the beginning of their career (without teaching experience). The negative interaction between gender and teaching experience indicated that the initial gender bias decreased with increasing years of teaching experience. The additional significant main effect of the continuous variable teaching experience ($b_{\text{exp}} = 0.02$) suggested that the fictive girl's grades improved by approximately 0.02 standard deviations per year of teaching experience. In the German male sample, only teaching experience ($b_{\text{exp}} = -0.01$) significantly influenced grading. Accordingly, with growing teaching experience, lower grades were awarded. While all of the other teachers showed a consistent bias pattern, the gender-neutral grading behavior of the German male teachers was exceptional. In the following analyses and figures that were aimed at gaining further information on gender bias effects as a function of teaching experience, I hence focus on the Swiss, Austrian, and German female teachers. Yet, it is important to always keep in mind that all that is reported in the following does not apply to the whole teacher sample. The reported gender bias effects are not generally valid and show only partial cross-border generalizability with the German male physics teachers demonstrating a divergent effect pattern.

In Figure 3.2, an interaction plot for the equated samples of Swiss, Austrian, and German female teachers based on Model 2 is depicted. To illustrate the moderating effect of teaching experience, grades were regressed on the mean teaching experience ($M = 17.76$, $SD = 11.41$) minus or plus one standard deviation and student gender, which was set to female or male.

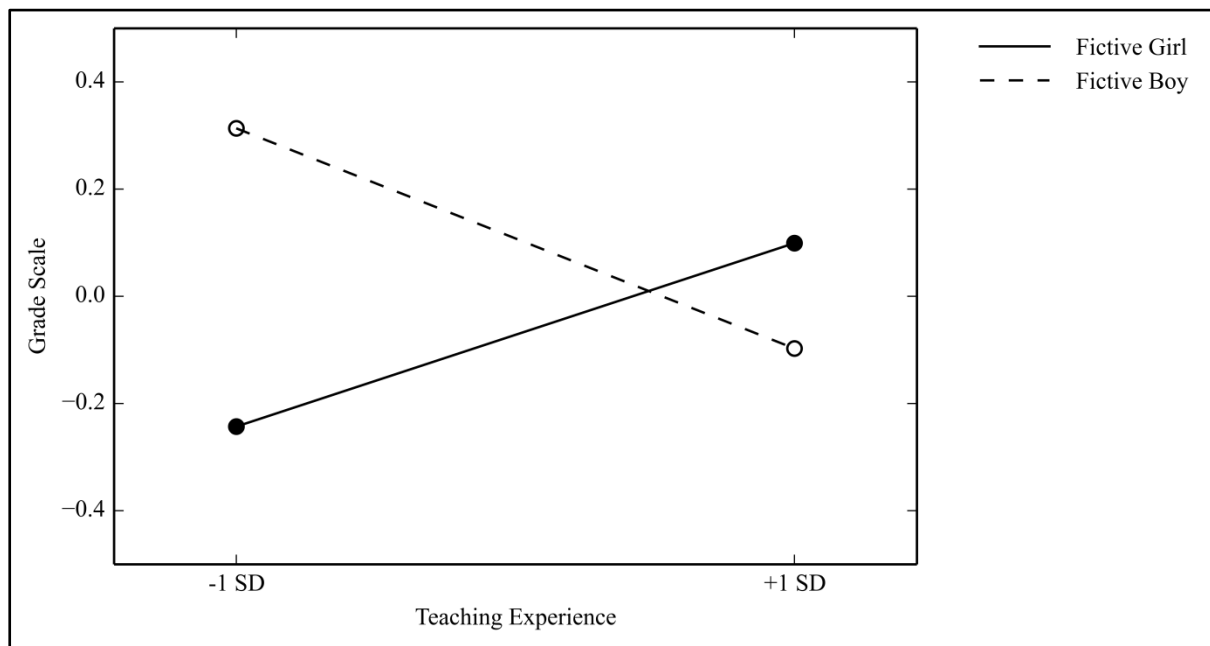


Figure 3.2. Interaction plot based on the equated samples of Swiss, Austrian, and German female teachers (Model 2). Grades were predicted by the mean teaching experience minus/plus one standard deviation and the fictive student's gender.

T-Tests on Binned Data

In keeping with Model 2, t-tests were applied only on the data from Swiss, Austrian, and German female physics teachers ($n = 378$). The comparisons between the mean grades that were awarded to a fictive female vs. male student in each of the nine five-year teaching experience bins revealed that after approximately ten years of teaching experience, the gender-specific grade discrepancy was not significant any more. Hence, the mean grade difference was $M_{\Delta} = 0.87$ ($t(51) = 3.61$, $p < .001$) in the first bin and $M_{\Delta} = 0.67$ ($t(54) = 2.40$, $p < .05$) in the second bin, compared to $M_{\Delta} = 0.15$ ($t(59) = 0.65$, $p = .52$) in the third bin (in all of the six other bins, all $ps \geq .34$). The problem of multiple testing (i.e., the nine t-tests) was considered negligible here taking into account the severity of the problem of even small bias effects in grading. Expressed in the country-specific unstandardized grade scales, a difference of $M_{\Delta} = 0.87$ on the z-standardized grade scale would correspond to about 0.7 Swiss grades, to about 0.9 Austrian grades, and to about 0.9 German grades. Figure 3.3 visualizes the relationship between teaching experience and grading based on both individual data points and binned data (i.e., the interpolation line) as a function of the fictive student's

gender. Importantly, Figure 3.3 additionally provides information about the number of teachers within each teaching experience bin in the form of histograms.

To conclude, the findings of this additional analysis suggested an interpretation of the interaction between gender and teaching experience in the sense that teaching experience removed the strong initial bias against girls but did not reverse it.

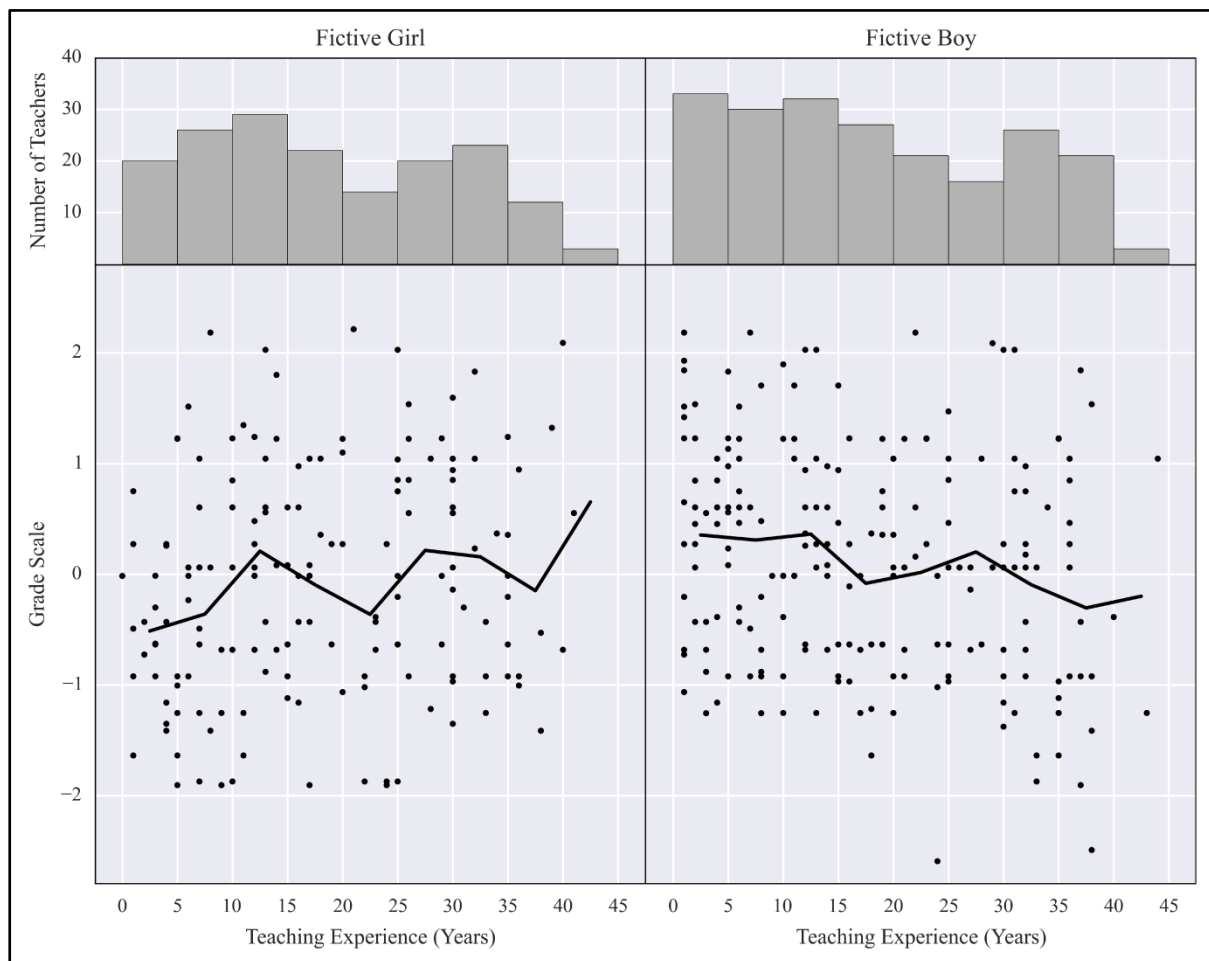


Figure 3.3. Scatter plots showing the relationship between teaching experience and grades for a fictive girl on the left side and a fictive boy on the right side, based on the samples of Swiss, Austrian, and German female physics teachers. The interpolation line connects the mean grades that were calculated within bins of five years of teaching experience. The histograms in the upper part of the figure visualize the number of teachers within each teaching experience bin.

Discussion

Gender Bias and Teaching Experience: Effects Exist, but not for All Teachers

This experimental online-study investigated gender bias and the role of teaching experience in grading a fictive student's answer to a conceptual test question in secondary school physics. Contrary to prior expectations, the overall sample displayed no generally valid pattern of bias effects. This finding indicated that bias effects do not generalize easily across contexts. Nevertheless, this study revealed the existence of a consistent cross-border pattern of gender bias effects that applies to Swiss, Austrian, and German female teachers. Overall, the inspection of the data regarding this international group of teachers suggested a consistent clear gap between girls' and boys' grades that was significant for teachers with up to about ten years of teaching experience and disappeared with increasing teaching experience. Yet, unexpectedly, German male physics teachers showed no gender bias effects at all. In the German male sample, performance of both female and male students was rated lower with years of teaching practice. The teacher samples from the three countries are highly comparable in terms of age and years of teaching experience. German physics teachers' training program and gender distribution also closely resemble the situation in Austria (while Swiss teachers' training program slightly differs from the other two countries and the gender distribution shows a more extreme preponderance of male teachers). There were no distortions in the sampling of the German male sample regarding teaching experience or age. In trying to find an explanation for the German male teachers' divergent pattern, differences between the German male sample and the other samples were examined in terms of the proportion of teachers teaching at rural vs. urban schools and in terms of the time spent with the survey. However, also these analyses revealed no irregularities. To additionally factor in the considerably larger sample size of the German male physics teachers, random subsamples including approximately 25% (i.e., $n \sim 100$) of the overall German male teachers sample were drawn and analyzed. In none of the five subsamples analyzed gender bias effects emerged, suggesting that the pattern of bias effects in the samples of Swiss, Austrian, and German female teachers was not merely an effect of distorted samples due to small sample sizes. Further research is required in order to detail the specifics of German male physics teachers that might relate to their differing, gender-neutral grading behavior. It remains to be

investigated how the patterns of female and male physics teachers in other countries compare with the two patterns revealed in this study, searching for regularities. Such research may help to understand in which contexts gender bias effects in physics grading can be expected and when they do not appear, providing important information for remediating interventions.

Focusing now on the pattern of bias effects found in the samples of Swiss, Austrian, and German female teachers, the moderating effect of teaching experience may partially explain the heterogeneity of existing findings on gender bias, where characteristics of the raters or judges were not taken into consideration (see Swim et al., 1989). Thus, the rater's experiences, with regard to the context of the judgment task, can play an important role in determining to what extent or whether or not a gender bias may arise. Future research that also closely examines the rater is needed to elucidate the process that underlies bias changing with experience (c.f. Kunda & Spencer, 2003; Kunda & Thagard, 1996). On the one hand, there is good reason to assume that the need to invoke stereotypes decreases to the extent that the perceived ambiguity of information and a high demand for cognitive resources in the judgment situation diminish with increasing experience. On the other hand, experience could also reduce gender bias by changing the stereotype itself via repeated exposure to individuals that challenge formerly held beliefs (c.f. Glock & Krolak-Schwerdt, 2013; Koenig & Eagly, 2014; Miller et al., 2014).

Student specialization in languages vs. science was not systematically considered in the grading process. This finding may indicate that student specialization, as compared to student gender, did not activate a shared social category that was used in the grading process examined in this study. In contrast to specialization, the fictive student's gender, however, seems to serve as a cue activating cognitive structures that systematically influence some teachers' decision making during grading.

Relevance for Physics Classrooms

The study investigated teachers' evaluations of a student's answer to one conceptual test question. The distinct gender bias effects found for Swiss, Austrian, and German female teachers who have been teaching for less than ten years underpin the importance of straightforward assessment criteria – which mimic the more elaborated cognitive schemata of experienced teachers – whenever student performance is evaluated and especially when ill-

defined conceptual problems are to be judged. By reducing the perceived ambiguity and cognitive overload of beginning teachers in the judgment situation, the need to draw on stereotypes and the resulting biases may be avoided. Moreover, these findings suggest that teacher education and teacher supervision should focus more strongly on supporting beginning teachers in monitoring their (socio-)cognitive processes when student achievement is evaluated.

In real classroom situations teachers get to know their students after a while, and this knowledge base may at least reduce the application of stereotypes (c.f. Kunda & Spencer, 2003). Nevertheless, a teacher's evaluations and grading at the beginning of the school year, which resemble the situation that was implemented in this study (i.e., little personal information), may lead to self-fulfilling prophecy (e.g., de Boer, Bosker, & van der Werf, 2010; Jussim & Eccles, 1992) or stereotype threat effects (e.g., Marchand & Taasoobshirazi, 2013; Nguyen & Ryan, 2008).

Limitations

In all of the three countries, the correlation between the teachers' years of teaching experience and the teachers' age was very high ($.86 \leq r \leq .90$). Consequently, with the cross-sectional design used in this study, it is difficult to determine whether teaching experience or the different socialization of the age cohorts influenced gender bias in grading. In trying to nevertheless estimate the relative impact of experience vs. age, an additional regression analysis was conducted for those Swiss, Austrian, and German female teachers with below average age and above average teaching experience ($n = 18$) and those teachers with above average age and below average teaching experience ($n = 30$). Albeit this analysis is based on very small sample sizes and the coefficients were not significant, a negative coefficient of the main effect of gender in the group of the younger but more experienced teachers and a positive coefficient in the group of the older but less experienced teachers suggest that not age but teaching experience determines the change in gender bias for Swiss, Austrian, and German female teachers. This conclusion, however, has to be underpinned by future research.

In this study, the application of gender-STEM stereotypes was not explicitly examined but deduced from the teachers' evaluation behavior and theoretical assumptions, since this study primarily aimed at describing physics teachers' gender bias in grading as a function of

teaching experience. Closely related, the domain specificity of the observed effect patterns was not addressed in this study. Now that there is evidence that gender bias effects in fact have to be considered in physics grading, further studies could add specific and detailed measures of (implicit) stereotype activation and application (see e.g., Glock & Kovacs, 2013; Nosek et al., 2009) that also allow for a differentiation between general gender-STEM stereotypes, on the one hand, and more specific gender-physics stereotypes, on the other hand. The generalizability of the observed patterns to other STEM fields also has to be investigated. Future work that is built upon the present results may further apply more comprehensive instruments to assess the teachers' evaluation of student performance which was measured only in the form of grades in this study.

The teachers' familiarity with the particular physics problem used in this study, the "skateboarder question", may be an alternative to teaching experience to explain the observed bias pattern. How often the teachers in the sample have come across a physics problem similar to the "skateboarder question" can be expected to depend on the length of their professional experience. Familiarity with the problem might have helped teachers in the study to interpret and evaluate the student answer by comparing it to mental representations of answers that different students have provided over the years. It can be argued, however, that familiarity is equivalent to the more and more efficient structuring of a physics problem's cognitive schema which is expected to proceed with teaching experience. Familiarity with single, frequently met problems (like the "skateboarder question") could hence be expected to inevitably accompany growing teaching practice. Future studies have to implement different and less familiar judgment situations to investigate familiarity with the test question as alternative explanation.

Conclusion

In the first decade of Swiss, Austrian, and German female physics teachers' careers, grading is affected by a gender bias that is in line with the common gender-STEM stereotypes. Gender bias disappears with increasing years of teaching practice. German male teachers, by contrast, display gender-neutral grading behavior. It remains to be clarified why this group of teachers behaves differently.

Despite the only partial generalizability of gender bias effects, even today gender bias in grading seems to represent a real problem in at least some physics classes. Since gender bias effects in grading should not appear at all, this finding has to be taken seriously. Ultimately, some girls' underperformance in physics may be an inevitable consequence of the social learning environment, while the existence of gender-STEM stereotypes may be an inevitable consequence of the girls' underperformance. Breaking this vicious circle by sensitizing student teachers and novice physics teachers to the problem of gender bias in grading and by providing straightforward strategies to assess student performance, could be one approach that would allow the gender gap in physics to be addressed.

References

- Babad, E. Y. (1985). Some correlates of teachers' expectancy bias. *American Educational Research Journal*, 22(2), 175–183. <http://doi.org/10.3102/00028312022002175>
- Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, 40(2), 191–202. <http://doi.org/10.1080/0013188980400207>
- Bartlett, S. F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.
- Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, 35(5), 463–482. [http://doi.org/10.1016/S0883-0355\(02\)00004-6](http://doi.org/10.1016/S0883-0355(02)00004-6)
- Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *A dual process model of impression formation* (pp. 1–36). Lawrence Erlbaum Associates, Inc.
- Carter, K., Sabers, D., Cushing, K., Pinnegar, S., & Berliner, D. C. (1987). Processing and using information about students: A study of expert, novice, and postulant teachers. *Teaching and Teacher Education*, 3(2), 147–157. [http://doi.org/10.1016/0742-051X\(87\)90015-1](http://doi.org/10.1016/0742-051X(87)90015-1)
- Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, 135(2), 218–261. <http://doi.org/10.1037/a0014412>
- Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, 66(3), 460–473.
- Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology*, 78(2), 141–146.
- De Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1), 168–179. <http://doi.org/10.1037/a0017289>
- Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten. *Zeitschrift für Pädagogische Psychologie*, 23(34), 187–195. <http://doi.org/10.1024/1010-0652.23.34.187>
- Eagly, A. H., & Koenig, A. M. (2008). Gender prejudice: On the risks of occupying incongruent roles. In E. Borgida & S. T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom* (pp. 63–81). Blackwell Publishing Ltd.
- Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, 8(3), 340–357. <http://doi.org/10.1177/1745691613484767>

- Eagly, A. H., Wood, W., & Diekmann, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- European Commission (2013). *She figures 2012: Gender in research and innovation*. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/she-figures-2012_en.pdf
- Farenga, S. J., & Joyce, B. A. (1999). Intentions of young students to enroll in science courses in the future: An examination of gender differences. *Science Education*, 83(1), 55–75.
- Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1 and 2 (4th ed.)* (pp. 915–981). New York, NY, US: McGraw-Hill.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, 23, 1–74.
- Glock, S., & Kovacs, C. (2013). Educational psychology: Using insights from implicit attitude measures. *Educational Psychology Review*, 25(4), 503–522. <http://doi.org/10.1007/s10648-013-9241-3>
- Glock, S., & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education*, 16(1), 111–127. <http://doi.org/10.1007/s11218-012-9197-z>
- Glock, S., & Krolak-Schwerdt, S. (2014). Stereotype activation versus application: How teachers process and judge information about students from ethnic minorities and with low socioeconomic background. *Social Psychology of Education*, 17(4), 589–607. <http://doi.org/10.1007/s11218-014-9266-6>
- Goddard Spear, M. (1984a). Sex bias in science teachers' ratings of work and pupil characteristics. *European Journal of Science Education*, 6(4), 369–377. <http://doi.org/10.1080/0140528840060407>
- Goddard Spear, M. (1984b). The biasing influence of pupil sex in a science marking exercise. *Research in Science & Technological Education*, 2(1), 55–60. <http://doi.org/10.1080/0263514840020107>
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109(1), 3–25. <http://doi.org/10.1037//0033-295X.109.1.3>
- Heller, K. A., Finsterwald, M., & Ziegler, A. (2010). Implicit theories of mathematics and physics teachers on gender-specific giftedness and motivation. In K. A. Heller (Ed.), *Munich studies of giftedness* (pp. 239–252). Berlin: LIT.

- Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement*, 22(3), 177–82.
- Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, 63(6), 947–961.
- Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9(2), 131–155.
- Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, 76(4), 761–780. <http://doi.org/10.1348/000709905X59961>
- Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, 107(3), 371–392. <http://doi.org/10.1037/a0037215>
- Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess: Der Lehrer als «flexibler Denker». *Zeitschrift für Pädagogische Psychologie*, 23(34), 175–186. <http://doi.org/10.1024/1010-0652.23.34.175>
- Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2012). Leistungsbeurteilungen von Schulkindern: Welche Rolle spielen Ziele und Expertise der Lehrkraft? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 44(3), 111–122. <http://doi.org/10.1026/0049-8637/a000062>
- Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, 129(4), 522–544. <http://doi.org/10.1037/0033-2909.129.4.522>
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103(2), 284–308.
- Leiner, D. J. (2014). *SoSciSurvey - oFb - der onlineFragebogen*. Retrieved from <https://www.soscisurvey.de>
- Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, 78(2), 75–95. <http://doi.org/10.1037/0022-0663.78.2.75>
- Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class, and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin*, 24(12), 1304–1318. <http://doi.org/10.1177/01461672982412005>
- Marchand, G. C., & Taasoobshirazi, G. (2013). Stereotype threat and women's performance in physics. *International Journal of Science Education*, 35(18), 3050–3061. <http://doi.org/10.1080/09500693.2012.683461>

- Miller, D. I., Eagly, A. H., & Linn, M. C. (2014). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000005>
- Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41), 16474–16479.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (Seventh Edition). Los Angeles, CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf
- Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93(6), 1314–1334.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math \neq me. *Journal of Personality and Social Psychology*, 83(1), 44–59. <http://doi.org/10.1037/0022-3514.83.1.44>
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1), 36–88. <http://doi.org/10.1080/10463280701489053>
- Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26), 10593–10597. <http://doi.org/10.1073/pnas.0809921106>
- Organisation for Economic Co-operation and Development (2011). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science* (Vol. 1). OECD Publishing.
- Palmer, D. J., Stough, L. M., Burdinski, Jr., T. K., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist*, 40(1), 13–25. http://doi.org/10.1207/s15326985ep4001_2
- Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, 50(4), 1262–1281. <http://doi.org/10.1037/a0035073>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333–343.

- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3), 743–762. <http://doi.org/10.1037/a0027627>
- Swim, J. K., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, 105(3), 409–429.
- UCLA: Statistical Consulting Group (2014, July 18). Mplus FAQ. How can I compute a chi-square test for nested models with the MLR or MLM estimators? Retrieved July 18, 2014, from http://www.ats.ucla.edu/stat/mplus/faq/s_b_chi2.htm

Acknowledgements

I wish to thank David Wintgens, Andreas Vaterlaus, Martin Hopf, Knut Neumann, Karsten Reckleben, Gerhard Röhner, Franz Kranzinger, and Silke Eckstein for their help in contacting physics teachers in Switzerland, Austria, and Germany; Elsbeth Stern, Andreas Lichtenberger, Clemens Wagner, Bahar Behzadi, André van der Graaff, Rolf Strassfeld, and Sebastian Seehars for their support; and, in particular, all of the physics teachers for their participation. I would also like to thank Bruno Rütsche and Peter Edelsbrunner for their help with technical and software issues.

4. The basic Mechanics Concept Test (bMCT): An Efficient Rasch-Scaled Multiple Choice Test of Fundamental Conceptual Understanding in Newton's Mechanics

Sarah I. Hofer, Ralph Schumacher, and Herbert Rubin

Solid assessment of understanding in Newton's mechanics is highly relevant both for physics classrooms and research. Several concept tests have been developed. What is still missing, however, is an efficient test that is adapted to the content taught to secondary school students and that can be validly applied as pre- and posttest to reflect learning progress. In this paper, we describe the development and evaluation of the basic Mechanics Concept Test (bMCT) that was designed to meet these requirements. In the context of test development, qualitative as well as quantitative methods including Rasch analyses were applied to a sample of $N = 239$ Swiss secondary school students. The final test's conformity to the Rasch model was confirmed with a sample of $N = 141$ students. We further ascertained the bMCT's applicability as change measure. Additionally, the criterion validity of the bMCT and the Force Concept Inventory (FCI) was compared in a sample of secondary school students ($N = 66$) and a sample of mechanical engineering students ($N = 21$). In both samples, the bMCT clearly outperformed the FCI in predicting actual student performance. The paper closes with a discussion on the bMCT's potential regarding physics education and research purposes.

Keywords: *Test construction; Newtonian mechanics; Rasch model; Conceptual understanding; Performance assessment*

Introduction

There is not much at school that troubles students as much as physics (see e.g., Beaton et al., 1996; Halloun & Hestenes, 1985; McDermott, 1984). Science and especially physics literacy, however, is becoming more and more relevant in an environment that is based on scientific and technological progress. Groundbreaking work has been done by Halloun and Hestenes (1985) and Hestenes, Wells, and Swackhamer (1992), who were the first to systematically investigate naïve concepts in mechanics, developed the well-known Force Concept Inventory (FCI; Hestenes et al., 1992), and initiated the idea of learning as conceptual change. According to this line of research, students have already built naïve concepts about scientific phenomena they constructed to explain everyday experiences when they enter the physics classroom for the first time. Those concepts often do not comply with scientifically accepted models that are taught at school and thus hamper learning (e.g., Duit, 2004; Hardy, Jonen, Möller, & Stern, 2006; Hestenes et al., 1992; Vosniadou, 1994). Familiar terms from everyday language that mean something completely different when used in the scientific context, further lead to confusion – just think about “force” and “work”, for instance (Brookes & Etkina, 2009; Rincke, 2011). With its many overlaps to everyday life, understanding Newtonian mechanics has proven particularly challenging (e.g., Halloun & Hestenes, 1985; Nieminen, Savinainen, & Viiri, 2010; Shtulman & Valcarcel, 2012). At the same time, however, Newton’s three axioms represent a core concept – how a body moves in response to forces acting upon it – that is taught in introductory lessons and provides the basis for later physics contents.

To be able to react to the students’ actual knowledge state (e.g., to their naïve concepts) and intervene appropriately, it is highly relevant to adequately assess the students’ conceptual understanding of Newton’s axioms. Accordingly, in the following sections, we first introduce the general idea of conceptual knowledge that is closely linked to the best known physics concept test, the FCI, which is addressed afterwards. We then outline shortcomings of existing tests, what finally sets the stage for the introduction of a new test of fundamental conceptual understanding in Newton’s mechanics, the basic Mechanics Concept Test (bMCT).

Conceptual Knowledge

Conceptual knowledge can be described as abstract and general knowledge of a domain's main principles and their connections (Carey, 2000; Schneider & Stern, 2010b). In the domain of physics, this kind of knowledge may exemplarily include understanding of “body movement in response to forces” or “momentum conservation” (see Halloun & Hestenes, 1985). The single principles that constitute conceptual knowledge are often referred to as concepts. Owing to its abstract nature, conceptual knowledge enables flexible problem solving that is not bound to specific contexts (see Hiebert, 1986). This kind of deep understanding, compared to problem-bound calculation routines or memorizing of formulae, can be considered the essential element of physics literacy.

Since prior knowledge determines the processing of new information (e.g., Carmichael & Hayes, 2001; Ohst, Fondu, Glogger, Nückles, & Renkl, 2014; Stern, 2009), understanding new concepts depends on the compatibility between the concept to be learnt and already existing conceptual knowledge that might encompass either partially or entirely wrong concepts. To enable learning in case of incompatibility, conceptual change has to take place (e.g., Posner, Strike, Hewson, & Gertzog, 1982; Schneider, Vamvakoussi, & Van Dooren, 2012). Several conditions have to hold to make conceptual change happen, including a dissatisfaction regarding the actual concepts as well as the perceived intelligibility, plausibility, and explanatory potential of the concept to be learnt (Posner et al., 1982). Specific information about the learners' conceptual knowledge state is hence essential for effective instruction to explicitly work with and on the students' existing concepts (c.f. Schneider & Stern, 2010a).

The Seminal Role of the FCI

A major step in understanding students' learning difficulties in physics was achieved when Hestenes, Wells, and Swackhamer (1992) presented their findings gathered with a new kind of assessment instrument, the Force Concept Inventory (FCI; Hestenes et al., 1992; Halloun, Hake, Mosca, & Hestenes, 1995). This test requires a choice between Newtonian and naïve concepts derived from everyday experience. Hestenes and colleagues (1992) demonstrated that even university students' beliefs about the physical world are mainly

derived from personal experience and to a large amount incompatible with Newtonian concepts. Since its publication the FCI has been successfully applied in a number of studies and raised awareness of both the existence and persistence of naïve concepts in diverse populations up to advanced physics students (e.g., Crouch & Mazur, 2001; Domelen & Heuvelen, 2002; Hake, 1998; Savinainen & Scott, 2002). By now, investigations of naïve concepts in a broad range of learning domains have been conducted (e.g., Hardy et al., 2006; Vosniadou, 1994) and further tests targeting heat and energy (Prince, Vigeant, & Nottis, 2012) or biology concepts (Klymkowsky & Garvin-Doxas, 2008) have been developed.

Shortcomings of Existing Concept Tests in Newton's Mechanics

Without undermining the FCI's seminal contribution to educational research, there are some shortcomings of the test that have to be considered. Hence, the inventory has been criticized to neither really measure a force concept nor the six conceptual dimensions (including kinematics, the first law, or the superposition principle) supposedly comprising it, as indicated by factor analyses (Henderson, 2002; Huffman & Heller, 1995; Saul, 1998; for a response of the FCI authors, see Hestenes & Halloun, 1995). Moreover, the test consists of 29 items implying long working time and high mental effort, both for those working on the test and for the one who has to analyze it. A second well-known test in mechanics is the more restricted Force and Motion Conceptual Evaluation (FMCE; Thornton & Sokoloff, 1998). It examines conceptual understanding of Newton's laws of force and motion (dynamics). With its 47 items, however, the FMCE is definitely not meant to provide an efficient, quick overview of students' understanding but detailed in-depth information about students' conceptual profiles in dynamics.

Both tests are designed to measure conceptual understanding in a diverse student population, from high school students to university students. Hence, they are not perfectly adapted to assess that kind of knowledge students at higher levels of secondary school are taught. Moreover, neither the FCI nor the FMCE were developed based on item response theory (IRT) or the Rasch model that enable the construction of strict measurement instruments whose conformity to certain quality criteria is statistically ascertained (see Bond & Fox, 2007; Hambleton & Jones, 1993; Lord, 1980).

Finally, concept tests are routinely applied as pre- and posttests to measure knowledge gains. Since gain scores are commonly used without checking for uniformity between pre- and posttest data (e.g., by means of IRT or the Rasch model), this approach has repeatedly been criticized (see Cronbach & Furby, 1970; Lohman, 1999). Analyses suggest that both the FCI (Planinic, Ivanjek, & Susac, 2010) and the FMCE (Ramlo, 2008) do not assess the same construct or latent dimension when applied as pretest (without formal instruction on the topic) vs. posttest (after instruction). Hence, simply comparing FCI or FMCE pre- and posttest data means comparing two different measures. Change then is assessed within uncertain frames of reference.

A New Instrument

The idea to assess conceptual knowledge about mechanics by means of a concept test as brought up by Hestenes, Wells, and Swackhamer (1992) and their FCI forms the basis for the new instrument introduced in this paper. Yet, what is missing in addition to existing tests is a user-friendly, short, and efficient mechanics concept test (*efficiency*) that is adapted to the content taught to secondary school students (*content validity*) and that can be validly applied both as pre- and posttest to reflect learning progress (*valid change measure*). To be able to meet these requirements, we drew on the Rasch model (Rasch, 1960) when constructing and evaluating our new instrument, the basic Mechanics Concept Test (bMCT).

Content validity was targeted by involving secondary school teaching experts in the bMCT's development process and explicitly adjusting the content to the subject material taught at the higher tracks of secondary school. To achieve *efficiency*, we aimed to choose a small number of items conforming to the Rasch model. In this model, item difficulty and a person's ability level are measured on the same invariant scale and simple sum scores can legitimately be used to unambiguously indicate a person's ability level due to exhaustive statistics. This characteristic of Rasch-scaled tests enables highly efficient testing with the test administrator only having to add up correct answers to obtain a valid estimation of a student's ability level (see Boone & Scantlebury, 2006). Finally, the instrument's *validity as change measure* was investigated by testing the fit of one uniform Rasch model on both pre- and posttest data. In addition, we determined the bMCT's reliability and compared the bMCT and the FCI in terms of their criterion validity regarding grades in a sample of secondary

school students and a sample of mechanical engineering students to gain further insights into the relative potential of the bMCT.

Method

In the following, we first provide an overview of the Rasch model. The R packages applied and the instrument's development are presented next. We then turn to the evaluation of the bMCT's final version. Hence, sample and participants are described before we finally outline the evaluation process itself.

The Rasch Model

The Rasch model is a psychometric model for binomial (dichotomous) data. The model assumes local stochastic independence and thus one-dimensionality. It further claims that every item has to equally contribute to the estimated ability level, implying equal item discrimination and, thus, the absence of an item discrimination parameter. Moreover, the Rasch model demands specific objectivity, stating independence of items when ability levels are compared and independence of ability levels when item difficulties are compared. If all these requirements are fulfilled, a test instrument can be considered as unequivocally measuring a single underlying dimension (see e.g., Bond & Fox, 2007; Strobl, 2010; Wright & Stone, 1979).

Its basic equation (see Equation 1) describes the difference between the ability of a specific person n , B_n , and the difficulty of a specific item i , D_i , by a logarithmic function that depends on the probability P_{ni} of person n to correctly solve item i :

$$B_n - D_i = \ln (P_{ni} / 1 - P_{ni}) \quad (1)$$

Hence, the person parameter (B_n) represents a person's ability level and the item parameter (D_i) constitutes an item's difficulty. As indicated by the subtraction on the left side of Equation 1, person and item parameters are measured on one scale. A specific person's probability P_{ni} to solve item i (right side of Equation 1) is dependent on the person's ability B_n and the item's difficulty D_i (left side of Equation 1). Consequently, if a specific person's ability B_n complies with a specific item's difficulty D_i , the person's probability to solve this

item is $P_{ni} = .50$ (see e.g., Boone, Staver, & Yale, 2014). There are several methods available to test both the global fit of the Rasch model on the data and the fit between the data and the model's assumptions of one-dimensionality (or local stochastic independence) and subgroup homogeneity (see Strobl, 2010).

Given the assumption of exhaustive statistics, sum scores represent all information on a person's ability level. Thus, item parameters can be estimated with person parameters omitted. Common estimation methods for item parameters are the conditional and the marginal Maximum Likelihood (ML) method. Person parameters can be obtained afterwards with the weighted ML method. Alternatively, simultaneous estimation of both parameters is conducted by means of the joint (also called unconditional) ML method (for more detailed information, see e.g., Linacre, 1998; Strobl, 2010).

R Packages

Throughout test development and evaluation, R (R Core Team, 2013) was used to examine fit to the Rasch model. We applied the packages eRm (Mair, Hatzinger, & Maier, 2013) and ltm (Rizopoulos, 2006) for fitting the Rasch model and comparing it to the two-parameter Birnbaum model, psychomix (Frick, Strobl, Leisch, & Zeileis, 2012) for testing Rasch mixture models, and nFactors (Raiche, 2011) for confirming one-dimensionality by means of factor analysis. The package sirt (Robitzsch, 2014) was used to calculate a marginal true score reliability.

Test Development


The bMCT was developed in a stepwise procedure with qualitative methods complementing quantitative item analyses.

Content validity. In the first phase of test development, the focus was on arriving at a set of items with high content validity. Initially, a group of physics and secondary school teaching experts as well as educational psychology experts constructed a pool of 22 multiple choice items targeting Newton's three axioms. We built multiple choice items and not single choice items (as in the FCI) to impede guessing. This approach also enabled us to survey

deep conceptual understanding, since students have to detect all correct and omit all incorrect answer alternatives.

8. A person is standing in a resting boat and tosses a big stone into the water behind the boat. Which of the following statements are true?

- ☐ The boat moves in the direction the stone was thrown.
- ☐ The stone displaces water and this is why the boat moves just slightly back and forth.
- ☐ If you let an inflated balloon whizz through the air, principally the same happens.
- ☐ The boat moves contrary to the direction the stone was thrown.



11. The following three balls are moving on a horizontal plane:

- Ball A is moving around a bend with a velocity of $1\frac{m}{s}$.
- Ball B starts with a velocity of $6\frac{m}{s}$ and then becomes slower and slower.
- Ball C is moving faster and faster.

Which of the following statements are true?

- ☐ A horizontal force is acting on ball A.
- ☐ A horizontal force is acting on ball B.
- ☐ A horizontal force is acting on ball C.

Figure 4.1. Two sample items of the basic Mechanics Concept Test (bMCT) translated into English. For item 8 (“Stone”), the last two answer alternatives are correct and for item 11 (“Balls”), all three answer alternatives are correct.

With these 22 items, each of the three axioms was broadly covered and every single item measured the underlying basic concept of how a body moves in response to forces acting

upon it. Problem contexts and wrong answer alternatives were inspired by teaching experiences and existing research (e.g., Halloun & Hestenes, 1985; Hestenes et al., 1992; Thornton & Sokoloff, 1998). While some items addressed only one axiom, others referred to a combination of them. Theoretically they intertwine and all refer to the same basic concept. We emphasize this one-dimensionality underlying all items, because we do not believe that, after instruction on all three axioms, it is possible to perfectly understand one of the three axioms without understanding the other two axioms. What is assumed to reflect the student's degree of understanding is how consistent a student can apply the basic concept of how a body moves in response to forces acting upon it across the three axioms and different problem contexts provided by the items. Because the bMCT should be particularly adapted to secondary school students, the problem contexts of the single items were less complex than in the FCI, for instance. So, for example, we refrained from having students evaluate parabolic trajectories that depend on a combination of forces and avoided problem contexts that require a lot of information provided a priori. We aimed to construct items as concise as possible without hampering their comprehensibility. Since the bMCT was intended to flexibly serve also as pretest, it was important to avoid specific terminology that is hard to understand without previous mechanics instruction (e.g., momentum, energy, gravitation). For illustration, Figure 4.1 shows two sample items that are included in the final version of the test.

In a first draft of the test, the multiple choice items were supplemented by requests to explain the choice or to draw a sketch. The test was given to several Swiss students with and without knowledge in mechanics attending the Gymnasium, the highest track in the Swiss educational system. Students' answering patterns, pictures, and comments were used to modify the test. This procedure was repeated several times until the items' intelligibility and appropriateness, also in terms of difficulty and item-scale correlation, were ascertained. Finally, interviews were conducted with a sample of $N = 6$ (3 girls) Gymnasium students between 13 and 17 years with and without knowledge in Newtonian mechanics. In the fashion of think-aloud protocols, the repeatedly modified set of 22 items was presented to the individual students without offering any answer alternatives. Their answers and considerations were recorded. The students' thoughts reflected the answer alternatives constructed and suggested only minor further modifications. When looking at item 8 ("Stone") that is presented in Figure 4.1, for instance, a student without prior knowledge and two students with prior knowledge suggested that the water displaced by the stone moves the

boat in the direction contrary to the direction the stone was thrown. They had a correct intuition for what is going to happen (boat moves in the opposite direction), but an incorrect explanation (waves). To be able to detect these faulty thoughts, we included a second correct answer alternative (“*If you let an inflated balloon whizz through the air, principally the same happens.*”) to assess deep understanding of the underlying abstract principle. One student expected that nothing is going to happen and two students expressed the correct idea. The students’ thoughts about the presented problem situation could hence be well mapped by the answer alternatives we had constructed, completed by the balloon analogy. We analyzed each item and the students’ thoughts about it in this way.

Efficiency. Having achieved a fixed set of good items, in the second phase of test development, the aim was to considerably reduce the number of items. Since the final test should satisfy the Rasch model to enable efficient usage, we checked the items for compliance with the Rasch model and excluded divergent items. When single items do not measure the same underlying dimension across subgroups, you speak of differential item functioning (DIF). For this last step, the 22 items were distributed to a sample of $N = 239$ (150 girls) Swiss Gymnasium students being on the average $M = 16.34$ ($SD = 1.40$, range 14–20) years old. Information on age, gender, mother tongue, potential areas of specialization at school, and prior knowledge on Newton’s mechanics was gathered in addition. If an item was answered correctly, that is, no wrong answer alternative and all correct answer alternatives checked, the item was scored $x_{ni} = 1$. Otherwise, the item was scored $x_{ni} = 0$. To detect differential item functioning (DIF), we used Andersen’s likelihood ratio test (Andersen, 1973) that examines the hypothesis that item parameter estimation does not vary between two subgroups and the analogous nonparametric T10-statistic (Ponocny, 2001) with different splitting criteria (bMCT mean, bMCT median, age, gender, mother tongue, specialization, prior knowledge). With the splitting criterion *gender*, for instance, item parameter estimates were compared between boys and girls. We expected no significant differences, given that the Rasch model holds. The splitting criterion *mean* implied that students scoring above average and students scoring below average on the bMCT were compared in terms of item parameters. When at least one of the tests indicated subgroup heterogeneity ($p < .05$), we continued with item-specific analyses. On the individual item level, the graphical model test with 95% confidence regions was conducted. This analysis estimates item difficulties separately for the two groups produced by the respective splitting criterion. The estimated item difficulties are plotted on two axes surrounded by confidence regions. An item’s

subgroup heterogeneity is then indicated by significant deviation from the diagonal. Furthermore, the item-specific Wald test that provides a significance test of the subgroup homogeneity assumption was inspected. Additionally, item-fit statistics were considered to reveal significant deviations in the answering patterns of individual items. We also fitted Rasch mixture models (Rost & von Davier, 1995) that search for latent classes within a sample indicated by maximally different item parameter estimations. When models with more than one class turned out to better fit the data than the regular (one class) Rasch model, we inspected the item difficulties estimated for the latent classes to find out what items especially differed between the detected latent classes. Whenever an item showed marked significant deviation or only slight discrepancies but on more than one statistic, it was excluded. In a stepwise procedure, we eliminated the least fitting item first, repeated all tests and eliminated the next item. We stopped as soon as the tests indicated no further violations of the Rasch model. The resulting test consisted of 12 items with three to ten answer alternatives each (item 4 “Train” and item 12 “Skaters” comprise two parts). For students’ ease of processing, the final items were ordered with increasing difficulty (see Appendix B).

In this way, we developed a concept test in mechanics that is adapted to the content taught to secondary school students (*content validity*). As a consequence of Rasch model conformity, the test moreover is user-friendly, short, and efficient (*efficiency*). The fit to the Rasch model, however, had to be confirmed with a new sample of students who took the final version of the bMCT. We also had to demonstrate the valid application of the bMCT both as pre- and posttest to reflect learning progress (*valid change measure*). To finalize the instrument’s evaluation, the bMCT’s reliability and criterion validity further had to be ascertained. Hence, the bMCT’s evaluation is described next.

Student Sample

The sample to finally evaluate the bMCT was taken from an ongoing research project that implements cognitively activating Newtonian mechanics instruction and compares it to conventional instruction in real physics classrooms. All $N = 141$ (69 girls) participants with a mean age of $M = 15.87$ ($SD = 1.10$, *range* 14-19) years were students from the Swiss Gymnasium who worked on the bMCT under supervision and without time pressure. In maximally 30 minutes all students managed to work through the 12-items test. The bMCT

was administered before instruction (pretest) and after instruction (posttest). Unless otherwise specified, we always refer to the students' bMCT posttest measure.

Evaluation Strategy

In the following, we first describe the steps taken to assess the final instrument's fit to the Rasch model substantiating its *efficiency*. The strategy to examine the bMCT's *validity as change measure* is delineated next. We, finally, briefly describe how we determined the bMCT's reliability and criterion validity.

Assessing the fit of the Rasch model. Pearson's χ^2 -goodness-of-fit (bootstrap) test assessed general model fit while the nonparametric T10-statistic was applied to gauge subgroup homogeneity. All nonparametric statistics were based on $n = 5000$ sampled matrices. We decided to examine DIF using gender, type of instruction, re-testing, and the medians of the bMCT, of age, and of intelligence as split-variables.

Gender. It was considered especially important that the bMCT measures boys and girls on the same scale. Gender-fair testing is essential in the context of performance assessment. Thus, we investigated DIF in terms of gender.

Type of instruction. A part of the sample ($n = 58$) received 18 lessons of introductory Newtonian mechanics instruction focusing on the conveyance of conceptual understanding. Methods such as metacognitive questions, self-explanations, holistic mental model confrontation, and inventing were implemented in this unit to help students grasp the meaning of the three axioms. We examined DIF between students having received this kind of instruction vs. conventional instruction ($n = 83$) in order to check whether different kinds of instruction differentially influence the probability to solve single items, what would ultimately change the meaning of the underlying basic mechanics concept that should be unambiguously measured independently of type of instruction. Generally, we made sure that there was no teaching to the test and that the content of the single items was not dealt with during instruction.

Re-testing. All students in the sample completed the bMCT repeatedly, as pre- and as posttest. Hence, their bMCT (posttest) measures represented measures that resulted after they had already worked on the test before instruction. Therefore, re-testing effects could be

expected that would be problematic in case that re-testing differentially influences the probability to solve single items. To be able to assess DIF when onetime vs. repeated testing are compared, we extended the sample by including parts of the old sample used to develop the test where the bMCT had been applied only once, after Newtonian mechanics had been instructed ($n = 108$). If item parameter estimation did not vary between the two samples, it could be inferred that re-testing had no effect on Rasch model conformity. This information would guarantee that the test always measures the same underlying mechanics concept no matter if it is applied as posttest in a pre-posttest study design or for educational purposes only once at the end of a school year, for instance.

bMCT. The median of the bMCT was used as split-variable to make sure that the instrument does not function differently for students having developed a rather good understanding of the underlying concept as compared to students with little or no understanding.

Age. We also checked for DIF regarding age differences. Hence, younger students could be expected to solve items differently as compared to older students who might already be further ahead in terms of their general cognitive development. The bMCT measure, however, should be directly related to conceptual understanding in mechanics with any other influences ruled out.

Intelligence. Finally, a student's intelligence level should not influence the items' difficulty ranking and the test's structure. Although intelligent students are expected to perform better on the bMCT measure than less intelligent students, differences in general intelligence should not lead to qualitative differences in how single items are solved. Intelligence was estimated by means of the set II score of Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1992; maximum score = 36).

To further examine subgroup homogeneity, Rasch mixture models with two and three classes were compared to the solution with only one class (the Rasch model). The nonparametric version of the Martin-Löf test was run to assess if different item-subsets all measure the same underlying dimension (see Verguts & De Boeck, 2000). This assumption could be violated, for instance, when effects of fatigue in the second half of the test systematically influence the measurement or when students manage to learn from the first items. Hence, the first half of the items was compared to the second half. The median of the

item-specific solution rates was used as another criterion to split the items. In addition, odd items were compared to even items. A maximum likelihood factor analysis with varimax rotation was performed to further check whether a one-factor-solution fits the data, while the global nonparametric T11-statistic was inspected to test for local stochastic independence. As a last check, the Rasch model was compared to the two-parameter Birnbaum model that includes a second parameter allowing the items to differ in discrimination. The less restrictive Birnbaum model should not fit the data better than the parsimonious Rasch model. Finally, item parameters (i.e., item difficulties) were estimated by means of the conditional ML method as implemented in the R package eRm (Mair, Hatzinger, & Maier, 2013).

The bMCT as change measure. To examine whether the bMCT measures the same latent dimension when applied as pretest (without prior knowledge) vs. posttest (with prior knowledge) the nonparametric T10-statistic was inspected to compare the item parameter estimation between pretest and posttest data. When evaluating the bMCT's *validity as change measure*, it was important to avoid dependencies in the data. Consequently, we ensured that every participant appeared only once in this sample, either with pretest or posttest data.

Reliability and criterion validity. Referring to Dimitrov (2003), we calculated population true score measures for our binary items calibrated with IRT to achieve true score reliability estimates. For the evaluation of the bMCT's criterion validity, we used two additional samples, a sample of secondary school students and a sample of mechanical engineering students, to investigate the bMCT's benefit in predicting actual student performance compared to the FCI. The additional secondary school students sample comprised $N = 66$ (38 girls) Swiss Gymnasium students from three physics classrooms with a mean age of $M = 16.53$ ($SD = 0.66$, *range* 15-18) years. The mechanical engineering students sample comprised $N = 21$ (2 girls) students in their first semester at the Swiss Federal Institute of Technology in Zurich. Both school students and university students had recently dealt with Newton's three axioms in their classes. All participants worked on the final bMCT and the German translation of the FCI by Gerdes and Schecker (1999) without time pressure. The order of the two tests was randomly interchanged so that half of the students in each school class and in the university students sample worked on the bMCT first while the other half worked on the FCI first. The number of items solved correctly (i.e., the sum score) for each test was used to predict the grade in Newton's mechanics in the school students sample

and the semester grade in Mechanics 1 (targeting Newton's mechanics) in the university students sample.

Results

In the following sections, we first present the findings concerning the fit of the bMCT data to the Rasch model. The evaluation of the bMCT as change measure is outlined next. This section closes with the results of the reliability analysis and the comparison between the bMCT and the FCI in terms of their criterion validity regarding grades.

The Fit of the Rasch Model

Regarding general model fit, the Pearson's χ^2 -goodness-of-fit (bootstrap) test suggested conformity of the data to the Rasch model ($p = .19$). All nonparametric T10-statistics indicated subgroup homogeneity with all $ps \geq .07$ (see Table 4.1).

Table 4.1

Nonparametric T10-Statistics with Different Split-Variables

Split-variable	Subgroup size		<i>p</i> -value
	n_1	n_2	
Gender	69	72	.07
Type of instruction	83	58	.35
Re-testing	108	141	.17
bMCT measure median	65	76	.11
Age median	48	93	.66
Intelligence (set II) median	61	64	.67

Note. The nonparametric T10-statistic gauges the homogeneity in the item difficulty parameter estimates between subgroups. The subgroups are determined by the six split-variables. All nonparametric statistics are based on $n = 5000$ sampled matrices. A non-significant p -value indicates no significant differences between subgroups in the item difficulty parameter estimation.

The comparison of Rasch mixture models with two and three classes to the solution with only one class (the Rasch model) additionally underpinned subgroup homogeneity with both the Bayesian information criterion (BIC) and the integrated classification likelihood (ICL) favoring the one-class-solution. The nonparametric version of the Martin-Löf test confirmed that all item-subsets tested against each other measured the same underlying dimension. Hence, the exact p -value was estimated at $p = .42$, when comparing the first half of the items to the second half. Using the median of the item-specific solution rates as split-criterion, an exact p -value of $p = .13$ resulted. The exact p -value was $p = .89$, when comparing odd items to even items.

Table 4.2

Item Difficulty D_i , Standard Error of D_i , 95% Confidence Interval of D_i , and Outfit Mean-Square (MSQ) for the 12 Items

Item	Item difficulty D_i	Standard error	95%-CI		Outfit MSQ
			LL	UL	
1. Water Glass	-2.12	0.17	-2.46	-1.79	0.92
2. Book	-1.24	0.14	-1.52	-0.97	0.97
3. Bus	-0.80	0.13	-1.07	-0.54	1.00
4. Train	-0.43	0.13	-0.69	-0.17	1.04
5. Walker	-0.25	0.13	-0.51	0.01	0.83
6. Wagon	-0.04	0.13	-0.30	0.22	1.12
7. Object Motion	0.05	0.13	-0.21	0.31	0.98
8. Stone	0.37	0.14	0.10	0.64	1.11
9. Inclined Plane	0.58	0.14	0.30	0.85	0.96
10. Motorcycle	0.62	0.14	0.34	0.90	0.87
11. Balls	1.33	0.17	1.00	1.65	0.78
12. Skaters	1.95	0.20	1.57	2.34	0.84

Note. CI = confidence interval; LL = lower limit, UL = upper limit. Item difficulty parameter D_i and its standard error estimated according to the Rasch model. Higher positive values indicate higher difficulty. Confidence intervals give an idea of the precision of the difficulty parameter estimation. Outfit MSQ is a fit statistic comparing expected with observed data patterns that is sensitive to outliers. Values around 1.00 (~ 0.50-1.50) indicate reasonable fit.

In line with these results, a maximum likelihood factor analysis with varimax rotation could substantiate the fit of the bMCT data to a one-factor-solution ($\chi^2 = 59.63$, $df = 54$, $p = .28$). Finally, also the global nonparametric T11-statistic suggested model fit when testing for local stochastic independence ($p = .29$). Referring to the BIC, the less restrictive Birnbaum model did not fit the data better than the parsimonious Rasch model. The results concerning the item parameter estimation are presented in Table 4.2. To base the item parameter estimation on a sufficiently large dataset, we applied the extended sample ($N = 249$) used to examine the effect of re-testing, since no influence of repeated vs. onetime testing on item parameter estimation could be observed (see Table 4.1).

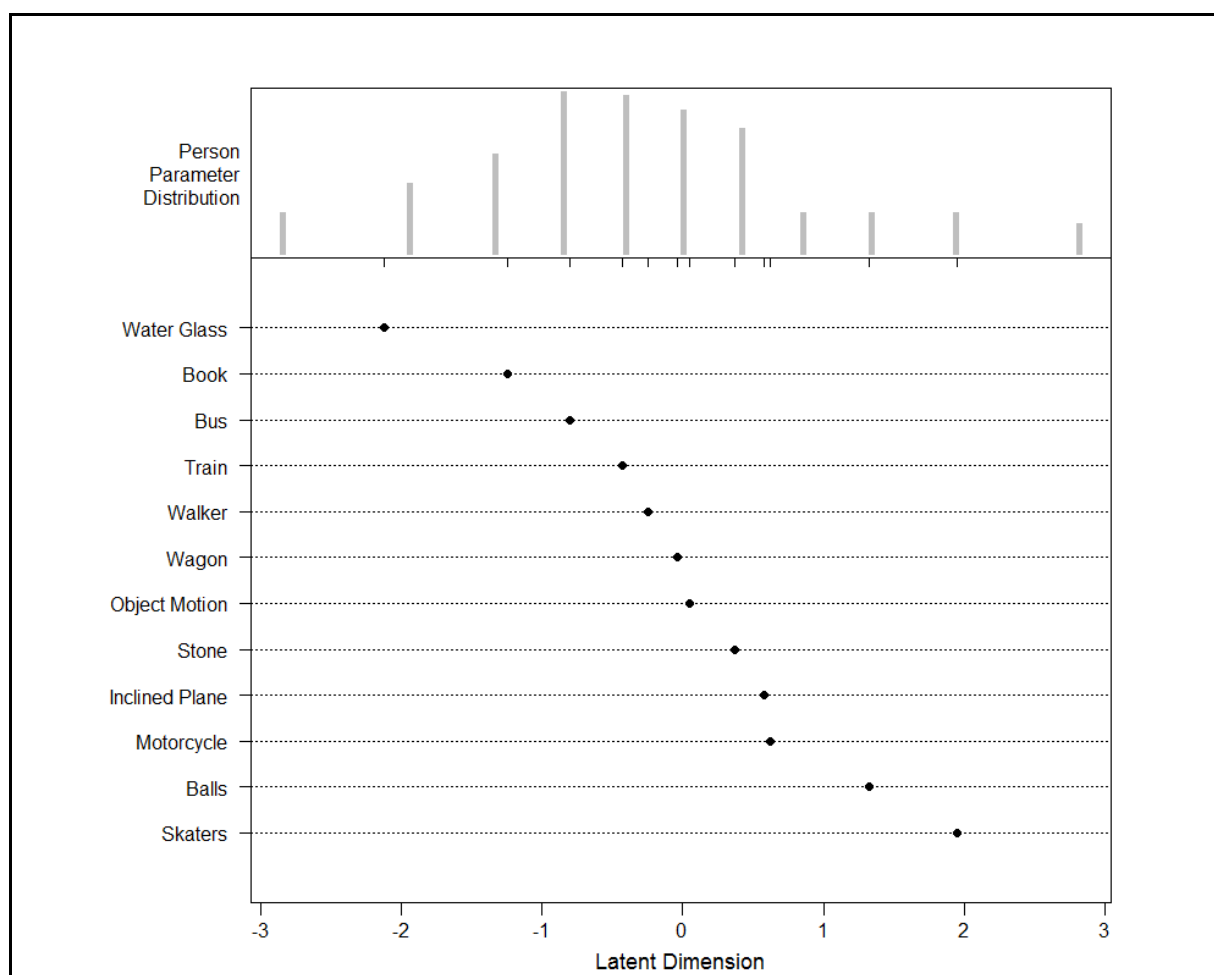


Figure 4.2. Wright Map on the 12 items of the bMCT and the extended student sample ($N = 249$). The upper panel provides the distribution of the students' person parameters and the lower panel depicts the location of each item's difficulty parameter along the same latent dimension. From the left to the right side of the map, person ability and item difficulty increase. The ticks on the horizontal axis of the upper panel reflect the locations of the items.

Figure 4.2 provides a Wright Map or person-item map (see e.g., Bond & Fox, 2007; Boone et al., 2014). This plot visualizes the location of each item's difficulty parameter (lower panel) together with the distribution of all person parameters (upper panel) along the latent dimension. The ordering of the items based on their locations on the latent dimension or their difficulty, respectively (see also Table 4.2), is in line with our expectations when constructing the items that differ in complexity. Comparing the item parameter distribution and the person parameter distribution, it becomes apparent that the items, which are reasonably spread across the latent dimension, cover large parts of the ability range of the students in the sample. As indicated by the mean and standard deviation of the item difficulty ($M = 0.19$, $SD = 0.92$) and the mean and standard deviation of the person ability ($M = -0.34$, $SD = 1.20$), with roughly 68% of all items being located between -0.73 and 1.11 on the latent dimension, the items concentrate on measuring the ability of a slightly more able student sample than the sample investigated, with roughly 68% of the students' ability parameters being located between -1.54 and 0.86 on the latent dimension. In the present student sample, however, the bMCT was administered without relevant external incentives. The mean ability of average secondary school students could hence be expected to slightly increase when the test is applied in a situation that is more relevant for the students, resulting in an increased fit between item difficulty and person ability. Moreover, the bMCT is intended to be used for research purposes, too. Interventions may be designed to enhance the students' performance. So, good differentiation in average to higher ability ranges is important. Nevertheless, also in the present sample, the items are well suited to measure the ability and differentiate between students in the average ability range where it is most relevant. At the same time, there are a few items located at both the lower and upper ends of the latent dimension to also allow some differentiation among especially low and high performing students. Yet, the more extreme regions of the latent dimension are less reliably measured. Since the aim was to provide an efficient test instrument with a small number of items, good coverage of the average ability range was considered more important than good coverage of the extreme regions.

The bMCT as Change Measure

The nonparametric T10-statistic was inspected to compare the item parameter estimation between pretest and posttest data with the resulting p -value ($p < .001$) disproving the notion of one uniform (prior knowledge-independent) Rasch model. The item-specific Wald test as

well as the graphical model test with 95% confidence regions indicated that especially for one item (item 2 “Book”), parameter estimations differed markedly ($p < .001$) between pretest and posttest. This item, dealing with normal force and its effect on a book lying on a table, was hard to solve correctly without instruction for all students, but easy after instruction. In their everyday lives, students usually do not consciously recognize phenomena related to normal force. Without being introduced to this kind of force in physics instruction, most students seem to have no idea about it. After this item was excluded, the nonparametric T10-statistic suggested conformity in the item parameter estimations between pretest and posttest ($p = .27$). With the exception of item 2 (“Book”), the test thus measures conceptual understanding of Newton’s three axioms on the same scale for pretest and posttest data, which allows a simple calculation of knowledge gain. Item 2 data, however, should only be used when the bMCT is applied after instruction. Consequently, changes between pretest and posttest should be assessed with the 11-items version of the bMCT. Following the already described procedure, Rasch model conformity could also be ascertained for the 11-items version of the bMCT.

Reliability and Criterion Validity of the bMCT

The marginal true score reliability of the bMCT was estimated at $\rho_{xx} = .69$ and at $\rho_{xx} = .67$ for the 11-items version and was hence acceptable to good. We evaluated the bMCT’s criterion validity in two additional samples. In the secondary school students sample, the bMCT measure and the FCI score correlated significantly with the grade in Newton’s mechanics ($r = .48$ and $r = .38$) and with each other ($r = .63$). The FCI score alone explained 14% of the variance in the grades ($p < .01$). When we included the bMCT measure into the regression, additional 10% of the variance in the grades could be explained ($p_{\text{change}} < .01$). In the regression with both predictors, only the bMCT measure significantly predicted grades. The bMCT measure alone explained 23% of the variance in the grades ($p < .001$). When we added the FCI score, no significant change in the prediction was achieved.

In the mechanical engineering students sample, the bMCT measure but not the FCI score correlated significantly with the semester grade in Mechanics 1 ($r = .56$ and $r = .26$). The two tests again correlated significantly with each other ($r = .67$). The FCI score alone explained 7% of the variance in the grades ($p = .26$). When we included the bMCT measure into the regression, additional 26% of the variance in the grades could be explained ($p_{\text{change}} < .05$). In

the regression with both predictors, only the bMCT measure significantly predicted grades. The bMCT measure alone explained 32% of the variance in the grades ($p < .05$). When we added the FCI score, no significant change in the prediction was achieved.

Discussion

In this paper we described the development and evaluation of a multiple choice test assessing fundamental conceptual understanding in Newton's mechanics. The way the instrument was constructed and evaluated enabled us to ascertain a user-friendly, short, and efficient mechanics concept test (*efficiency*) that is adapted to the content taught to secondary school students (*content validity*) and that can be validly applied both as pre- and posttest to reflect learning progress (*valid change measure*). Moreover, the test was sufficiently reliable. We could show that the bMCT significantly predicted mechanics grades not only in a sample of secondary school students, but also in a sample of mechanical engineering students. Consequently, the bMCT turned out to be a valuable predictor of university students' mechanics understanding, too, although it was not explicitly designed for this advanced student group. The FCI, by contrast, did not significantly contribute to predict inter-individual grade differences in either sample. Moreover, the bMCT and the FCI correlated substantially with each other, further indicating criterion validity. The correlation between the bMCT and the FCI was higher than the correlation between the bMCT and grades. This finding suggests that differences in conceptual understanding, as measured by the two tests, are indeed reflected in grade differences but cannot fully explain the inter-individual variation in the grades (about 67% variance unexplained). In the following, we discuss the bMCT's implementation in educational practice and research before considering some limitations.

The bMCT's Potential for Physics Instruction

The bMCT can be easily implemented in the physics classroom. Test instructions are short and readily understandable for secondary school students. The test can be processed in approximately 20 minutes and analyzed in one minute per test by simply checking answer alternatives and summing up all items solved without mistakes (all correct answer alternatives and no wrong answer alternative marked). The distribution of the item

parameters that represents each item's difficulty enables a differentiated measurement of a person's ability in the average achievement range with most items covering this area. At the same time, however, there are both two easily solvable, encouraging items and two particularly difficult items that allow assessment at the top end and prevent ceiling effects. We ascertained that the instrument assesses the same underlying ability for girls and boys as well as different age and intelligence groups. The test, moreover, unambiguously measures conceptual understanding in Newton's mechanics independent of the quality of physics instruction and whether the test has been solved only once or repeatedly. Consequently, whenever fair, exact, and efficient assessment of understanding in Newton's mechanics is required, the bMCT may be considered. The test could accordingly complement summative assessments but also be highly valuable in the context of formative assessment (e.g., Centre for Educational Research and Innovation, 2005; Wiliam, 2010), where efficiency is especially important.

The bMCT's Potential for Research

The bMCT constitutes a valuable instrument for assessing effects of interventions in the context of Newton's mechanics. The instrument can be applied to compare different instructional approaches guaranteeing a fair measurement without qualitative differences in item processing and conceivability. This fair measurement results from Rasch conformity and has explicitly been tested when comparing bMCT item parameter estimates for students having received cognitively activating instruction vs. conventional instruction. A major advantage of the bMCT compared to existing instruments is its confirmed applicability as change measure. With one item excluded, the resulting 11-items version of the bMCT measures the same underlying concept independent of the students' prior knowledge. Hence, the test can be validly applied both as pre- and posttest to reflect learning progress. The bMCT has already been successfully implemented as pre- and posttest in a classroom intervention study (Hofer, Stern, Rubin, & Schumacher, 2015). If the whole 12-items test is to be used for pretest to posttest comparisons, racking and stacking procedures are recommended to calibrate pre- and posttest data (see Wright, 2003).

Referring to the comparison between the bMCT's and the FCI's criterion validity, the bMCT may be preferred to the FCI when inter-individual differences in the performance of secondary school and university students are to be assessed. The bMCT's potential for

application at the university level, however, has to be further investigated with larger and more diverse samples.

Limitations

The bMCT does not provide differentiated information about students' conceptual profiles representing, for instance, the individual hierarchy of naïve conceptions, intermediate conceptions, and scientific conceptions about Newton's mechanics. Such kinds of conclusions are explicitly intended in a number of other instruments in a variety of domains originating primarily from conceptual change research (e.g., Garvin-Doxas & Klymkowsky, 2008; Hardy et al., 2006). In the bMCT's development, we focused more on efficiently assessing conceptual understanding than on the measurement of conceptual profiles. The latter would have required the analysis of single answer alternatives instead of only looking at the items solved correctly. Nevertheless, the selection of some wrong answer alternatives and the omission of particular correct answer alternatives in the bMCT are associated with the activation of certain misconceptions in the specific problem context provided by the items. For instance, a student who does not select the second answer alternative of item 11 ("Balls"; see Figure 4.1), may have activated the misconception that there is no force necessary to slow down something. Hence, it is possible to additionally describe the students' performance on the single answer alternative level to examine in which problem contexts what kinds of misconceptions show up. However, we do not believe that all wrong answer alternatives can be classified as one of several systematically occurring misconceptions. Some wrong answer alternatives simply signify that a student does not possess the conceptual understanding necessary to solve the respective item correctly.

Albeit our results seem quite stable, sample sizes were only moderate and further research is needed to substantiate the bMCT's quality. The sample, moreover, consisted of secondary school students from the highest track of the Swiss educational system, the Gymnasium. Hence, all results reported here only refer to this population. It might thus be indicated to examine additional secondary school student populations from other school types and countries.

To conclude, the bMCT proved to enable efficient and at the same time rigorous measurement of fundamental conceptual understanding in Newton's mechanics. This new instrument might be fruitfully applied both in physics classrooms and educational research.

References

- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38(1), 123–140.
- Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates Publishers.
- Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, 90(2), 253–269. <http://doi.org/10.1002/sce.20106>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer Netherlands.
- Brookes, D., & Etkina, E. (2009). “Force,” ontology, and language. *Physical Review Special Topics - Physics Education Research*, 5(1), 1–13. <http://doi.org/10.1103/PhysRevSTPER.5.010110>
- Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, 21(1), 13–19. [http://doi.org/10.1016/S0193-3973\(99\)00046-5](http://doi.org/10.1016/S0193-3973(99)00046-5)
- Carmichael, C. A., & Hayes, B. K. (2001). Prior knowledge and exemplar encoding in children's concept acquisition. *Child Development*, 72(4), 1071–1090. <http://doi.org/10.1111/1467-8624.00335>
- Centre for Educational Research and Innovation (2005). *Formative Assessment: Improving Learning in Secondary Classrooms*. Paris: OECD Publishing.
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change” - Or should we? *Psychological Bulletin*, 74(1), 68–80.
- Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970–977.
- Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, 27(6), 440–458. <http://doi.org/10.1177/0146621603258786>
- Domelen, D. J. V., & Heuvelen, A. V. (2002). The effects of a concept-construction lab course on FCI performance. *American Journal of Physics*, 70(7), 779–780. <http://doi.org/10.1119/1.1377284>
- Duit, R. (2004). Schülervorstellungen und Lernen von Physik. *IPN Kiel*.
- Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, 48(7), 1–25.

- Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE-Life Sciences Education*, 7(2), 227–233. <http://doi.org/10.1187/cbe.07-08-0063>
- Gerdes, J., & Schecker, H. (1999). Der Force Concept Inventory. *Der mathematische und naturwissenschaftliche Unterricht*, 52(5), 283–288.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. <http://doi.org/10.1119/1.18809>
- Halloun, I. A., Hake, R. R., Mosca, E. P., & Hestenes, D. (1995). Force Concept Inventory (Revised, 1995).
- Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53(11), 1056–1065. http://doi.org/10.1007/978-3-642-20072-4_12
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.
- Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking." *Journal of Educational Psychology*, 98, 307–326. <http://doi.org/10.1037/0022-0663.98.2.307>
- Henderson, C. (2002). Common concerns about the Force Concept Inventory. *The Physics Teacher*, 40, 542–547. <http://doi.org/10.1119/1.1534822>
- Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to Huffman and Heller. *The Physics Teacher*, (33), 502–506.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141–158. <http://doi.org/10.1119/1.2343497>
- Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.
- Hofer, S. I., Stern, E., Rubin, H., & Schumacher, R. (2015). *Fostering conceptual understanding with cognitively activating instruction in physics classrooms: Evidence for general effects and special benefits for high potential students*. Manuscript in preparation.
- Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher*, 33, 138–143. <http://doi.org/10.1119/1.2344171>
- Klymkowsky, M. W., & Garvin-Doxas, K. (2008). Recognizing student misconceptions through Ed's tools and the Biology Concept Inventory. *PLoS Biol*, 6(1), e3. <http://doi.org/10.1371/journal.pbio.0060003>
- Linacre, J. M. (1998). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement*, 3(4), 382–405.

- Lohman, D. F. (1999). Minding our p's and q's: On finding relationships between learning and intelligence. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 55–76). Washington, DC: American Psychological Association.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Mair, P., Hatzinger, R., & Maier, M. J. (2013). *eRm: Extended Rasch Modeling. R package version 0.15-3*. Retrieved from <http://CRAN.R-project.org/package=eRm>
- McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, 37(7), 24–32. <http://doi.org/10.1063/1.2916318>
- Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics - Physics Education Research*, 6(2), 1–12. <http://doi.org/10.1103/PhysRevSTPER.6.020109>
- Ohst, A., Fondu, B. M. E., Glogger, I., Nückles, M., & Renkl, A. (2014). Preparing learners with partly incorrect intuitive prior knowledge for learning. *Frontiers in Psychology*, 5(JUL). <http://doi.org/10.3389/fpsyg.2014.00664>
- Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics - Physics Education Research*, 6(1), 1–11. <http://doi.org/10.1103/PhysRevSTPER.6.010103>
- Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, 66(3), 437–459. <http://doi.org/10.1007/BF02294444>
- Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2), 211–227.
- Prince, M., Vigeant, M., & Nottis, K. (2012). Development of the heat and energy concept inventory: Preliminary results on the prevalence and persistence of engineering students' misconceptions. *Journal of Engineering Education*, 101(3), 412–438. <http://doi.org/10.1002/j.2168-9830.2012.tb00056.x>
- Raiche, G. (2011). nFactors: Parallel analysis and non graphical solutions to the Cattell scree test (Version 2.3.3). Retrieved from <http://CRAN.R-project.org/package=nFactors>
- Ramlo, S. (2008). Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics*, 76(9), 882–886. <http://doi.org/10.1119/1.2952440>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.
- Raven, J. C., Raven, J., & Court, J. H. (1992). *Raven's Progressive Matrices und Vocabulary Scales. Teil 4 Advanced Progressive Matrices*. (S. Bulheller & H. Häcker, Trans.). Frankfurt: Swets & Zeitlinger.

- R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rinke, K. (2011). It's rather like learning a language: Development of talk and conceptual understanding in mechanics lessons. *International Journal of Science Education*, 33(2), 229–258. <http://doi.org/10.1080/09500691003615343>
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25.
- Robitzsch, A. (2014). Package “sirt” (Version 0.45-23). Retrieved from <http://www-star.stat.ac.uk/cran/web/packages/sirt/sirt.pdf>
- Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 257–268). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4612-4230-7_14
- Saul, J. M. (1998). *Beyond problem solving: Evaluating introductory physics courses through the hidden curriculum* (Dissertation). University of Maryland, College Park.
- Savinainen, A., & Scott, P. (2002). Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education*, 37(1), 53–58. <http://doi.org/10.1088/0031-9120/37/1/307>
- Schneider, M., & Stern, E. (2010a). The cognitive perspective on learning: Ten cornerstone findings. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 69–90). Paris: OECD Publishing.
- Schneider, M., & Stern, E. (2010b). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology*, 46(1), 178–192. <http://doi.org/10.1037/a0016701>
- Schneider, M., Vamvakoussi, X., & Van Dooren, W. (2012). Conceptual change. In *Encyclopedia of the Sciences of Learning* (pp. 735–738). Retrieved from http://link.springer.com/content/pdf/10.1007/978-1-4419-1428-6_352.pdf
- Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2), 209–215. <http://doi.org/10.1016/j.cognition.2012.04.005>
- Stern, E. (2009). The development of mathematical competencies: Sources of individual differences and their developmental trajectories. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Evidence from the Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC)* (pp. 221–236). Mahwah, NJ: Erlbaum.
- Strobl, C. (2010). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* (Vol. 2). München, Mering: Rainer Hampp Verlag.

- Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, 66(4), 338–352. <http://doi.org/10.1119/1.18863>
- Verguts, T., & De Boeck, P. (2000). A note on the Martin-Löf test for unidimensionality. *Methods of Psychological Research Online*, 5, 77–82.
- Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, 4(1), 45–69.
- Wiliam, D. (2010). The role of formative assessment in effective learning environments. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning. Using research to inspire practice* (pp. 135–159). Paris: OECD Publishing.
- Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2 or pre-test vs. post-test. *Rasch Measurement Transactions*, 17(1), 905–906.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Acknowledgements

We wish to thank all the physics teachers (particularly Lars Fleig, Patrick Spengler, and Erich Schurtenberger) who provided their classes in the process of test construction and evaluation; and all the students for working on the tests. We would also like to thank Michal Berkowitz for making contact with a sample of mechanical engineering students.

5. Fostering Conceptual Understanding with Cognitively Activating Instruction in Physics Classrooms: Evidence for General Effects and Special Benefits for High Potential Students

Sarah I. Hofer, Elsbeth Stern, Herbert Rubin, and Ralph Schumacher

Secondary school physics instruction is confronted with the students' deficient physics literacy in general and some girls' underachievement in particular. In this study, we investigate the potential of cognitively activating (CogAct) physics instruction that is focused on conceptual understanding to address these two issues. While positive effects of single CogAct instructional elements have already been confirmed, little is known about the effectiveness of a whole CogAct teaching unit implemented in physics classrooms. Four teachers participated with two classes each. They taught one of their classes based on an 18-lessons unit of CogAct instruction, while the other class was instructed as always with instructional time and content matched. Across three points of measurement, we gathered measures of conceptual understanding, quantitative problem solving, and motivation of $N = 172$ (92 girls) Swiss secondary school students. The results of multiple regression analyses showed that CogAct instruction was superior to conventional instruction in terms of both performance measures. CogAct students, however, required a conceptual scaffold at the quantitative problem solving posttest in order to outperform conventional students at follow-up problem solving. The advantages of CogAct instruction were not reflected in any of the motivational variables in the overall sample. Additional latent profile analyses revealed that underachieving girls, high achieving boys, and, particularly, high achieving girls profited considerably from CogAct instruction regarding performance and motivation. We discuss the findings of the present study in the light of the potential of transformed instruction and assessment to promote effective and gender-fair physics learning.

Keywords: *Physics/Science instruction; Cognitively activating; Conceptual understanding; Underachievement; Gender*

Introduction

A substantial part of students lacks knowledge of basic physics concepts even after several years of physics instruction at school. This has been extensively demonstrated in Newtonian mechanics, a field that forms the basis for later physics contents (e.g., Beaton et al., 1996; Halloun & Hestenes, 1985; Hestenes, Wells, & Swackhamer, 1992; McDermott, 1984; Nieminen, Savinainen, & Viiri, 2010). Such findings are alarming in light of the crucial role of physics literacy in a society that is based on technological and scientific progress. While it is important to support all students to understand basic physics concepts, one group of students requires special attention: intelligent students who fail to realize their potential in physics. We know from prior research that such physics underachievers systematically occur among girls (Hofer & Stern, 2015). Hence, to speak of effective physics instruction, it is not sufficient to demonstrate mean learning advantages compared to conventional instruction. It is necessary to ascertain gender-fair physics instruction to reduce the gender gap in physics. To be able to investigate the potential of instruction to promote physics literacy and tackle underachievement, we have to go beyond short interventions or laboratory experiments. Based on instructional principles and elements tested in controlled experiments, prolonged classroom interventions with high ecological validity have to be implemented and evaluated.

Learning research has developed a number of such instructional elements that are described as cognitively activating since they encourage deep and focused processing and active construction of conceptual knowledge (see Berthold & Renkl, 2010; Schneider & Stern, 2010a). These include, for instance, confronting students with models or situations that are incompatible with their naïve concepts (e.g., Gadgil, Nokes-Malach, & Chi, 2012; Sanchez, Garcia-Rodicio, & Acuna, 2009), instructing students to generate self-explanations (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, De Leeuw, Chiu, & LaVancher, 1994; DeLeeuw & Chi, 2003; Schworm & Renkl, 2007; Ziegler & Stern, 2014), or prompting students to generate answers to metacognitive questions that deal with the topics at hand (e.g., Mevarech & Fridkin, 2006; Mevarech & Kramarski, 2003). Instruction of teachers with a cognitive constructivist orientation indeed benefits all students (Baumert et al., 2010; Staub & Stern, 2002). Cognitively activating (CogAct) instruction has the potential to even activate the highly intelligent but underachieving girls, since it aims at the construction of conceptual knowledge and does not require the students to memorize information and formulae without understanding (see e.g., Boaler, 1997; Zohar, 2006; Zohar & Sela, 2003). We assume that

these girls fail to memorize information and formulae without being able to make sense of them (c.f. Boaler, 1997; Hofer & Stern, 2015; Zohar & Sela, 2003). Conventional physics instruction often does not focus on imparting the conceptual knowledge that underlies formulae and problem solving routines (see Langer Tesfaye & White, 2012; Seidel et al., 2006; Taconis, 1995). Moreover, there is evidence that conceptual understanding does not only manifest in conceptual knowledge but is also necessary to develop flexible and transferable quantitative problem solving skills (e.g., Dixon & Moore, 1996; Hardiman, Dufresne, & Mestre, 1989; Leppävirta, Kettunen, & Sihvola, 2011). Instruction that fosters conceptual understanding may hence also promote quantitative problem solving.

In general, the ultimate usefulness and effectiveness of any curriculum program depends on the teachers' reception and implementation of the program in their classrooms (Newcombe et al., 2009; Remillard, Herbel-Eisenmann, & Lloyd, 2011). Adding to the already high demands of regular classroom interaction, teachers and students may be unable to make use of a new approach to learning that requires the adaptation to unfamiliar methods and to a different structuring of the content focusing on conceptual understanding, for instance. Promising innovations also have to demonstrate their effectiveness when integrated in everyday school life. Yet, instruction proven successful in controlled environments often is not tested in real classroom situations, implemented by teachers. It may also be the case that studies describing the failed implementation of instructional innovations in real classrooms are scarcely published. The study on implementing formative assessment by Yin and colleagues (2008) is one of the few exceptions providing valuable insights into the conditions that influence the (un-)successful implementation of learning research interventions. Teachers, for instance, omitted important parts of the formative assessment intervention (e.g., giving feedback to the students) or did not appreciate formative assessment as a means to improve learning and teaching (see also Furtak et al., 2008). A recently published special issue on implementation science significantly furthers our understanding of how to translate theory to practice (P. K. Murphy & Cromley, 2015). In line with Yin and colleagues' (2008) experiences, investigating, measuring, and increasing implementation fidelity is identified as crucial component that can shed more light on the effects detected or not detected, respectively (e.g., Greene, 2015; Star et al., 2015). Hence, teachers have to understand what they implement and why. They have to get the idea underlying the whole intervention (e.g., Cervetti, Kulikowich, & Bravo, 2015; Festas et al., 2015; Harris, Graham, & Adkins, 2015). Referring to the theory-practice gap when it comes to bringing insights from conceptual

change research into the classrooms, Duit and Treagust (2003) emphasize the importance of promoting the assimilation of new ideas into teachers' routines to take effect.

To conclude, it is far from certain that instructional elements proven successful in controlled experimental studies also result in superior learning when implemented by real teachers during instruction in real classroom situations. This study hence examined the effects of an 18-lessons unit of CogAct physics instruction taught by teachers during their regular physics classes. To scaffold implementation, we offered an elaborated CogAct teaching unit structuring and exemplifying the intervention. CogAct instruction was compared to conventional instruction that covered the same content of basic Newtonian mechanics. As dependent variables, we analyzed conceptual understanding, conventional quantitative problem solving, and motivational variables. We could show that CogAct physics instruction did not only result in mean learning advantages regarding students' conceptual understanding and quantitative problem solving but also boosted the performance of high potential students and tackled underachievement in real physics classrooms.

In the following sections, we present the theoretical basis this study is developed on. Hence, the importance of conceptual knowledge for learning in physics is outlined first. Next, we address the general effectiveness of instruction focusing on conceptual understanding regarding both conceptual and procedural knowledge development, before we summarize CogAct instructional elements that have proven successful in supporting conceptual understanding. Motivational variables that may be of interest in the context of CogAct physics learning are briefly described. We then characterize underachievement in the domain of physics and, finally, introduce the present study.

Conceptual Knowledge and Conceptual Learning in Physics

Educational psychologists all agree on the seminal role of conceptual knowledge and conceptual learning in the STEM fields (science, technology, engineering, mathematics; e.g., Carey, 2000; Organisation for Economic Co-operation and Development, 2006). Conceptual knowledge enables flexible, context-independent problem solving (see Hiebert, 1986) which is considered a key element of physics literacy (McDermott, 1984; Resnick, 2010).

“Conservation of energy”, “the uncertainty principle”, or “body movement in response to forces” are all examples of more or less complex concepts in the field of physics.

Yet, there is also broad consensus that conceptual learning is considerably constrained by already existing knowledge, consulted to make sense of the new information that is to be learnt. If possible, new information is assimilated. New information that is not compatible with the existing knowledge is omitted or forgotten in the long run. The knowledge structures that already exist prior to instruction have been constructed based on repeated everyday experiences attempting to explain and understand the world’s phenomena. They have been activated many times long before the first physics lesson. Extensive research, particularly in the field of Newtonian mechanics, has shown that these already existing knowledge structures are in fact highly persistent and extremely difficult to modify in formal instruction (e.g., Carey, 2000; Muller, Sharma, & Reimann, 2008; Ohlsson, 2013; Shtulman & Valcarcel, 2012; Smith III, diSessa, & Roschelle, 1994; Vosniadou, 1994; Vosniadou, Ioannides, Dimitrakopoulou, & Papademetriou, 2001).

As a consequence, a significant part of learners leaves school without having managed to understand basic physics concepts (e.g., Beaton et al., 1996; Halloun & Hestenes, 1985; Hestenes et al., 1992; McDermott, 1984; Nieminen et al., 2010). It is the responsibility of physics instruction to offer a learning environment that attunes to the learners’ conceptual knowledge state, demonstrates the limits of their existing knowledge, and introduces new concepts in a way that constructively builds on the learners’ existing knowledge (see Hardy, Jonen, Möller, & Stern, 2006; Hewson & Hewson, 1983; Mason, 2001; Posner, Strike, Hewson, & Gertzog, 1982; Schneider & Stern, 2010a; Schneider, Vamvakoussi, & Van Dooren, 2012; Smith III et al., 1994).

Effectiveness of Conceptual Instruction: Conceptual and Procedural Knowledge

Is there reason to assume that emphasizing conceptual understanding results in learning advantages both on conceptual and procedural measures? The relationship between conceptual understanding and procedural knowledge has been extensively studied in the field of mathematics. The results, however, are heterogeneous and vary considerably across studies

and the specific content domains (see Rittle-Johnson & Siegler, 1998). The majority of studies deal with basic mathematics like counting or single-digit addition that are relevant at early stages in the child's development. Dixon and Moore (1996), targeting higher level mathematics problem solving that requires proportional reasoning (a temperature mixture task) in older children, found that understanding the underlying concept was indispensable but not sufficient to develop procedures to mathematically implement the concept. Similar results were achieved for fractional arithmetic (Byrnes & Wasik, 1991). Concentrating teaching on mathematical algorithms even turned out to negatively affect both conceptual understanding and problem solving as compared to constructivist instruction that included inventing own solution procedures (Kamii & Dominick, 1997).

Considerably less research has been conducted on the connection between conceptual and procedural knowledge in the domain of physics. Nevertheless, the existing results resemble those obtained in the context of higher level mathematics. Focusing on classical mechanics, Hardiman, Dufresne, and Mestre (1989) hence found that better problem solving was correlated with learners' ability to categorize problems based on deep structure similarity. Successful problem solvers turned out to use principles or concepts to analyze the problems in the categorization task. The authors accordingly consider principles to play a crucial role in structuring and building procedural knowledge. Examining experts' and novices' physics problem solving, Chi, Feltovich, and Glaser (1981) could show in their seminal study that both the ability to think about physics problems in terms of underlying principles or concepts (vs. focusing on superficial features of the problem statement) and knowledge linking procedures with concepts in the form of explicit production rules may develop with expertise. Experts are reported to first qualitatively analyze a given physics problem in terms of the underlying principles or concepts. The concept or principle then determines the general procedure applicable to solve the problem. To conclude, again, conceptual understanding resulting in the development of conceptual knowledge seems to guide the activation of appropriate procedural knowledge. Heyworth (1999) obtained similar results in chemistry and likewise emphasized the seminal role of conceptual understanding in the development of conceptual and procedural knowledge. In their meta-analysis on teaching science problem solving, Taconis, Ferguson-Hessler, and Broekkamp (2001) concluded that instruction focusing on the underlying schemata seems to be effective, while practicing problem solving was of little importance. In a more recent study, advanced high school students had to compare their own knowledge structure to their teachers' knowledge

structure. Students who generated conceptual reflections in the process improved more on a concept relatedness rating task than students who generated procedural reflections (Sarwar & Trumpower, 2015). Hake (1998) explicitly compared traditional instruction with interactive engagement methods focusing on conceptual understanding at the high school, college, and university level. He reported interactive engagement methods to outperform traditional instruction both on a conceptual measure and a more quantitative problem solving test (see also Thacker, Kim, Trefz, & Lea, 1994). A comprehensive meta-analysis, finally, could show that students in undergraduate STEM courses based on active learning outperformed students having received traditional lecturing on concept inventories and, although less distinctively, also on course examinations that tend to focus on recall and quantitative problem solving (Freeman et al., 2014).

Approaching the question how conceptual understanding and procedural knowledge relate from the opposite perspective, there is also evidence that procedural knowledge is not enough to generate conceptual understanding. Byun and Lee (2014), for instance, found no correlation between the number of physics problems solved and conceptual understanding. In a study by Leppävirta, Kettunen, and Sihvola (2011), there was no effect of exposure to complex multistep problem exercises on university students' conceptual understanding of electromagnetics, whereas their procedural knowledge improved significantly. In line with this result, there is evidence that procedures are likely to improve and develop when explicitly trained. Algorithms can be applied to solve similar problems even without conceptual understanding (c.f. Gabel, Sherwood, & Enochs, 1984; Kamii & Dominick, 1998; Redish, Saul, & Steinberg, 1998). To conclude, instruction that focuses on practicing procedures or problem solving routines can produce successful domain-specific problem solvers. Existing research suggests, however, that such kind of instruction is unlikely to support the construction of conceptual knowledge. Instruction that focuses on conceptual understanding, by contrast, can be expected to promote conceptual knowledge that in turn guides the construction of procedural knowledge. The literature overview further indicates that neither conceptual nor procedural knowledge alone automatically lead to the development of both kinds of knowledge. Even after instruction emphasizing conceptual understanding, learners still have to actively build the necessary procedures. From a physics classroom perspective, there is no point in withholding any kind of important information from the students requiring them to generate it themselves. Therefore, from a practical point of view, physics instruction in real classrooms should never waive teaching procedures.

When thinking about fostering conceptual understanding and practicing procedures in the school context, the important question is where to place the emphasis. Existing research suggests that a focus on conceptual understanding may indeed pay off.

Finally, irrespective of its undisputed theoretical value, it has to be said that the practical distinction between conceptual and procedural knowledge is by far less clear-cut than the use of these two terms might suggest. This fuzziness is reflected in the large variation in research outcomes on the same topic depending on the conceptual and procedural measures implemented and evidenced by the poor convergent and divergent validity of the measures used to differentiate conceptual and procedural knowledge (Schneider & Stern, 2010b). Being aware of these limitations, the present study does not claim to offer conclusions about the development of pure conceptual or pure procedural knowledge. Under the premise of high ecological validity, we examine the effects of an instruction that is intended to foster conceptual understanding on a concept inventory-like questionnaire and on quantitative problem solving. These measures are probably the two kinds of assessments relevant for physics instruction at school that represent conceptual knowledge, or procedural knowledge, respectively, most unambiguously. While the former particularly targets conceptual knowledge, the latter, which implies the application of problem solving routines like setting up and solving equations, is considered to require procedural knowledge (e.g., Anderson & Schunn, 2000) – although not exclusively (c.f. Hake, 1998). We refer to these two kinds of assessments when we speak of conceptual and procedural knowledge in the context of this study.

Cognitively Activating Instructional Methods

As became clear, conceptual understanding can be expected to play a prominent role in the development of broad physics literacy. So how can instruction facilitate conceptual understanding? In the last two decades, research on learning and instruction has established different instructional methods and principles that are characterized as cognitively activating (see Berthold & Renkl, 2010; Schneider & Stern, 2010a). They stimulate learners to actively reorganize, augment, and construct conceptual knowledge and promote conceptual change. Some of the most promising methods and principles that are also integrated into the CogAct teaching unit are briefly summarized in the following sections.

Introducing New Topics with “Unexplainable” Phenomena

Knowledge construction and reorganization starts with the learner’s insight that a given problem cannot be solved referring to already acquired concepts (see Chinn & Brewer, 1993; Sanchez et al., 2009; Sinatra & Pintrich, 2003). To involve students into active knowledge construction, they have to be confronted with interesting phenomena they cannot explain, revealing the limits of their existing knowledge.

Inventing

Learning can be promoted by instructing students to invent a concept before the scientific concept is introduced (e.g., Schwartz, Chase, Oppezzo, & Chin, 2011; Schwartz & Martin, 2004). Learners are presented with several cases illustrating a specific underlying concept (e.g., linear graphs with different slopes) and instructed to discover the concept (e.g., to invent a common index that can be used to describe the slopes of these linear graphs). After the completion of the invention task, they receive the scientific explanation. This instructional method requires students to actively deal with a given problem and activate relevant prior knowledge, helping them to understand and process subsequent instruction (see also Kapur, 2008).

Self-Explanations

Self-explanations are explanations that are constructed for and addressed to oneself in order to clarify and rethink concepts. There is broad evidence that prompting self-explanations by specific questions is an effective way of enhancing students’ understanding (e.g., Chi et al., 1994; Rittle-Johnson, 2006; Schworm & Renkl, 2007; Siegler, 2002). Self-explanation prompts ask students to deliberate central points of the learning content. In addition to improving students’ understanding, repeatedly prompting self-explanations also encourages students to consider self-explanations as a generally effective learning strategy.

Holistic Mental Model Confrontation

When learners have to change their ideas about the relational structure of complex concepts, confronting their flawed mental models with an expert's conceptual model has proven effective (Gadgil et al., 2012). Learners are instructed to describe relevant differences between laypersons' and experts' models. In this way, common misconceptions are challenged.

Metacognitive Questions

Metacognitive questions prompt students to reflect their state of knowledge and their learning progress. There is broad evidence for the effectiveness of metacognitive questions (Mevarech & Fridkin, 2006; Mevarech & Kramarski, 2003; White & Frederiksen, 1998; Zepeda, Elizabeth, Ronevich, & Nokes-Malach, 2015; Zohar & Peled, 2008). Moreover, repeated training with metacognitive questions can improve the students' self-regulatory learning strategies.

Mental Tools

Knowledge transfer can be supported by mental tools like diagrams and graphs that direct the learner's attention to the abstract common elements of superficially different tasks (Hardy, Schneider, Jonen, Stern, & Möller, 2005; Novick & Hmelo, 1994). The active construction of linear graphs, for instance, turned out to have positive effects on the students' ability to transfer their knowledge between tasks with different contents (Stern, Aprea, & Ebner, 2003).

Connecting Concepts to Real-World Applications

Flexible knowledge is characterized by multiple connections between abstract concepts and concrete examples instantiating them (see Bereiter, 1997; King, 1994; Schneider & Stern, 2010a). When concepts and their concrete applications are represented together, the retrieval of relevant information in transfer situations is facilitated (Roth, Van Eijck, Reis, & Hsu, 2008). In addition, if a concept is connected to several concrete applications of the concept,

comparisons help to recognize abstract similarities between these different applications that reflect the underlying concept. Learners are hence supported to grasp the underlying concept (e.g., Gentner, Loewenstein, & Thompson, 2003; Loewenstein, Thompson, & Gentner, 2003).

Motivation in Cognitively Activating Physics Learning

In addition to foster conceptual and procedural knowledge, improving the students' attitudes towards physics, their inclination to engage in physics, and their perceived competence to succeed in physics have to be considered important objectives of physics instruction in light of the acute shortage of (especially female) students who opt for a career in the STEM fields (e.g., Nicholls, Wolfe, Besterfield-Sacre, Shuman, & Larpkiattaworn, 2007; Osborne, Simon, & Collins, 2003). CogAct physics instruction that focuses on conceptual understanding can be expected to increase the students' interest in physics, physics self-concept, and use of efficient learning strategies, as well as decrease the students' learning amotivation and physics anxiety as compared to conventional instruction¹. While interest may be raised by emphasizing and working on the conceptual explanations for various (formerly "unexplainable") phenomena, real-world applications, and formulae in active communication (Kiemer, Gröschner, Pehmer, & Seidel, 2015), the active role of the students involving student authorship in the knowledge acquisition process (e.g., by means of self-explanation prompts, inventing, or supporting self-monitoring of learning with metacognitive questions) may enhance the students' self-concept (Jansen, Scherer, & Schroeders, 2015; Zepeda et al., 2015). The repeated exposure to and application of methods like self-explanations, metacognitive questioning, or mental tools can be expected to stimulate students to use efficient learning strategies more often on their own (Vosniadou et al., 2001; White & Frederiksen, 1998; Zepeda et al., 2015). Likewise, the cognitively demanding instruction that requires each individual student to actively work on her/his knowledge structures and the explicit connections to real-world applications may reduce boredom, frustration, and, thus, learning amotivation (Deslauriers, Schelew, & Wieman, 2011; Hart, 1996; Kiemer et al., 2015; Zepeda et al., 2015). Physics anxiety may be

¹ For the sake of simplicity, we use the term 'motivation' throughout this article to summarize variables that could also be referred to as metacognitive, meta-strategic, affective, or self-belief variables, for instance.

addressed by focusing less on quantitative procedures and calculation, but more on conceptual understanding and discussion, and less on the transformation of given information, but more on the joint construction of comprehensible knowledge and self-regulated activities (Kesici, Baloglu, & Deniz, 2011; Kostova, 2015).

The Problem of Physics Underachievement

In particular, low interest and self-concept in physics have been identified as characteristics of one especially problematic group of learners: intelligent students who fail to realize their intellectual potential in physics (Hofer & Stern, 2015). The authors applied multiple group latent profile analysis on a sample of secondary school students to identify gender-specific student profiles. These profiles were based on the similarity on the two indicator variables intellectual potential and physics grades. The systematic co-occurrence of high intellectual potential and low physics grades defined physics underachievement. A profile of clear physics underachievers was detected for girls but not for boys. They exhibited an intellectual potential similar to physics high achievers and at the same time the worst physics grades of all of the profiles. These underachieving girls, who accounted for 29% of all female students, showed average school performance in subjects other than physics. This result indicates that they struggled particularly with physics classes, which are often focused on practicing formulae application and memorizing and leave little room for working on the underlying concepts (Langer Tesfaye & White, 2012; Seidel et al., 2006; Taconis, 1995; Zohar & Sela, 2003). In line with this, the finding that the underachieving girls were least interested in physics as compared to all of the other student profiles suggests that conventional physics classes (encompassing instruction and assessment) may in fact discourage and repel these girls (c.f. Hart, 1996; Kahle & Lakes, 1983; P. Murphy & Whitelegg, 2006a; Zohar, 2006; Zohar & Sela, 2003). Moreover, the underachieving girls seem to believe that they are not capable of doing physics, despite their high intellectual potential (Hofer & Stern, 2015; Jansen, Schroeders, & Lüdtke, 2014). There is evidence that conventional physics instruction tends to be oriented more towards boys than girls, with boys receiving more attention, being challenged by more demanding questions, and being expected to be more talented than girls (e.g., Andersson, 2010; Heller, Finsterwald, & Ziegler, 2010; Taasobshirazi & Carr, 2008). CogAct instruction that explicitly contrasts with conventional physics instruction by emphasizing deep conceptual understanding and the individual's role

in constructing knowledge can be assumed to disable some of the hypothesized negative effects conventional physics instruction might have on these intelligent girls (c.f. Häussler & Hoffmann, 2002; Hoffmann, 2002; Hulleman & Harackiewicz, 2009; Lorenzo, Crouch, & Mazur, 2006; Siegle, Rubenstein, & Mitchell, 2014; Zohar, 2006; Zohar & Sela, 2003).

The Present Study

Over the last decades, educational psychology has accumulated extensive knowledge about which instructional methods may be especially beneficial for learning. At the same time, secondary school physics instruction acutely suffers from a failure to ascertain physics literacy and prevent some girls' underachievement. This study hence broadly examined the potential of combining CogAct instructional methods within one expansive teaching unit that is implemented by physics teachers in their physics classrooms. Importantly, we did not aim at investigating the specific contributions of single instructional methods, but the effect of one CogAct teaching unit that targets conceptual understanding. Uniquely, we analyzed the potential of a CogAct teaching unit by looking at effects on a conceptual transfer measure as well as on quantitative problem solving performance, on the students' physics motivation, and, finally, on physics underachievement. The instruction covered introductory Newtonian mechanics, a topic that has turned out to be particularly susceptible to misconceptions and at the same time forms the basis for later physics contents. We investigated the effects of an 18-lessons unit of CogAct physics instruction taught by physics teachers during their regular secondary school physics classes. CogAct instruction was compared to conventional instruction that covered the same content of basic Newtonian mechanics in the same time span. Each of the four participating teachers taught one class according to the CogAct teaching unit and one class as always. We analyzed student data across three measurement points (pre, post, and follow-up) to be able to answer the following four research questions.

Research Question 1: General Effectiveness

Is a CogAct teaching unit beneficial for all students in terms of a conceptual transfer measure when compared to conventional instruction in physics classrooms? We analyzed immediate (post) effects and long-term (follow-up) effects after approximately three months.

Based on the existing research showing beneficial effects of specific CogAct methods and of general instructional approaches emphasizing the learner's role in actively constructing conceptual knowledge (e.g., interactive engagement, active learning), we hypothesized positive effects of CogAct instruction on deep conceptual understanding that should manifest in a conceptual transfer measure.

To assess the general effectiveness of CogAct instruction, we further had to examine whether a focus on conceptual understanding, as it is the case in the CogAct teaching unit, also proves beneficial for the acquisition of procedural problem solving skills. Again, we were interested in immediate (post) effects and long-term (follow-up) effects after approximately three months. As became apparent in the literature overview on the effectiveness of conceptual instruction, there was reason to expect CogAct instruction to promote conceptual knowledge that in turn guides the construction of procedural knowledge resulting in better quantitative problem solving performance than conventional instruction that can be assumed to put less emphasize on conceptual understanding.

Research Question 2: Accessing Procedures via Concepts

It can be assumed that students in general are used to solve quantitative problems applying problem solving routines triggered by certain cues in the problem context (c.f. Gick, 1986). Students with CogAct instruction spend less time with practicing such problem solving routines and rather learn procedures as abstract formulations or instantiations of the respective concept. They might have started to build knowledge structures that link the necessary procedures with the concepts in the form of explicit production rules, as described by Chi, Feltovich, and Glaser (1981). Hence, CogAct learners may have to access problem solving procedures by first activating the respective concepts. Consequently, because we did not expect students to automatically think about underlying concepts when trying to solve quantitative physics problems, half of the students within each classroom received an additional scaffold together with each quantitative problem. The scaffold simply prompted the students to think about the physics terms and principles that have to be considered in this problem before they start calculating. We hypothesized that this scaffold should be more beneficial for CogAct students, helping them to access the problem solving procedure via their conceptual knowledge, than for conventional learners who may have developed a less

elaborated conceptual knowledge base that less strongly connects to quantitative problem solving procedures.

Research Question 3: Impact on Motivation

Does CogAct instruction increase the students' interest in physics, physics self-concept, and use of efficient learning strategies, as well as decrease the students' learning amotivation and physics anxiety as compared to conventional physics instruction? The findings available so far in fact suggest positive effects of CogAct instruction on student motivation.

Research Question 4: Impact on Physics Underachievement

Given that physics underachievement is defined by the systematic co-occurrence of high intellectual potential and low physics grades, can physics instruction based on the CogAct teaching unit tackle underachievement as compared to conventional physics instruction? In particular, we examined the existence of physics underachievement and compared the probability to be an underachiever in both conditions. There were three possible outcomes that could be expected: First, the student profile of physics underachievers might not exist after CogAct instruction but after conventional instruction. Based on the findings by Hofer and Stern (2015), physics underachievers can be assumed to exist after conventional instruction. This outcome would indicate that CogAct instruction can in fact prevent physics underachievement. Second, underachievers might exist in both conditions, but the probability to be an underachiever might be reduced for CogAct instruction compared to conventional instruction. This outcome would still suggest beneficial effects of CogAct instruction on physics underachievement, although underachievement would not be prevented completely. Third, the existence of underachievers as well as the probability to be an underachiever might be independent of condition. This outcome, however, would allow two possible conclusions: First, underachievers did not profit from CogAct instruction. Second, potential effects of the CogAct intervention might not be reflected in the students' physics grades that were autonomously assigned by the participating teachers. To be able to clarify this point if necessary, we planned to additionally investigate the influence of CogAct instruction on the study measures within the student profiles defined by intelligence and physics grades (i.e., within the underachievers or high achievers, for instance). The physics underachievers'

conceptual understanding, quantitative problem solving, and physics motivation could thus be compared between CogAct and conventional instruction. Thereby, potential effects of CogAct instruction on physics underachievement reflected in measures other than physics grades could be detected.

We hypothesized that CogAct instruction may have the potential to tackle physics underachievement – if not in terms of physics grades, maybe in terms of the study measures conceptual understanding, quantitative problem solving, and motivation.

Method

In the following sections, we first explain the study's general design and then describe the student sample and the procedure. After the teacher training is outlined, the CogAct teaching unit is introduced. Finally, we present the measures implemented and the statistical methods that were applied to analyze the data corresponding to the four research questions.

Design

This quasi-experimental study applies a control-group design. Four Swiss higher secondary school physics teachers (three males, one female) volunteered to participate with two parallel classes each so that they could teach one class according to the CogAct instruction and one class as always (see Figure 5.1). Thus, the specific influence of each teacher was controlled by having each teacher instruct both conditions (CogAct and conventional). The parallel classes did not only share the physics teacher but also learnt in a highly comparable environment. On the individual student level, age, gender, specialization on non-STEM or STEM subjects, intelligence, and prior conceptual understanding were considered to potentially play a role in predicting the study performance measures within and across classrooms. These five variables were included as control variables. The design of the present study thus allowed for controlling important sources of additional variance that might otherwise distort potential effects of the CogAct instruction.

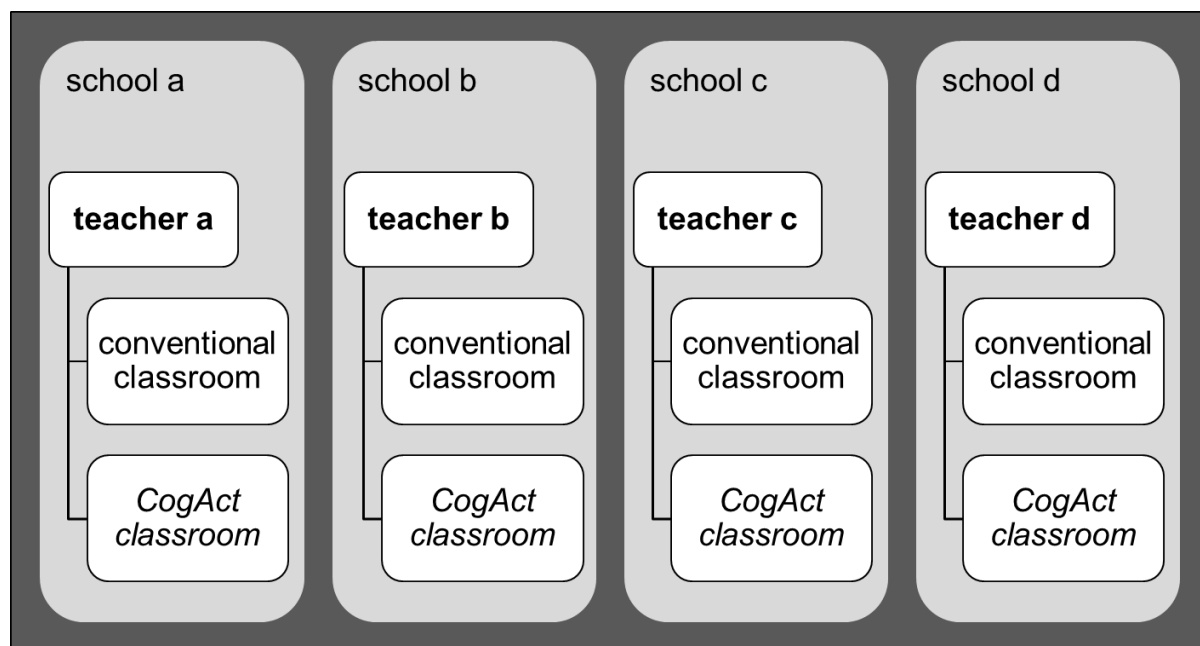


Figure 5.1. Basic design of the study.

Altogether, four classes hence received the CogAct instruction and four classes received conventional instruction (i.e., the control group). Which of each teacher's two classes was instructed according to the CogAct teaching unit was selected randomly. The four physics teachers were acquired by directly contacting teachers from schools that have already cooperated with our research institute or teachers who had expressed their interest in our work before. These physics teachers could be considered motivated and committed regarding the study as well as regarding their teaching in general. While two of the teachers had taught physics for less than ten years, the other two teachers were more experienced physics teachers. Student data were gathered at three measurement points (pre, post, and follow-up).

Student Sample

The four participating physics teachers' eight classes constituted a total of $N = 172$ (92 females) students. Eighty-seven of the students (48 females) received CogAct instruction. The students had a mean age of $M = 15.96$ years ($SD = 0.96$ years). They attended four different Swiss higher secondary schools (Gymnasien) located in three cantons of Switzerland with always two classes from one school taught by the same teacher. All students

and their parents were informed about the study and the parents' written consent was obtained.

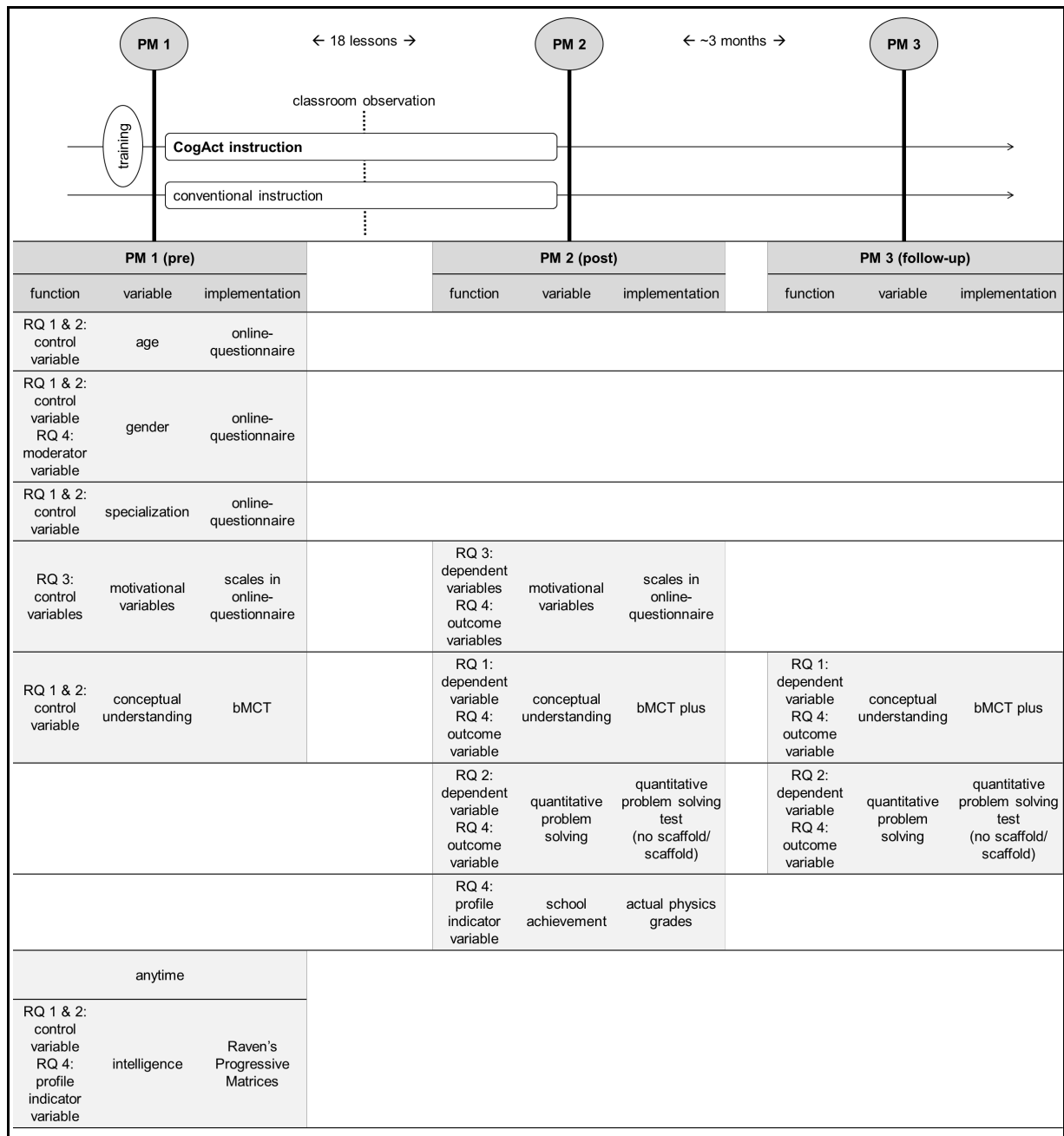


Figure 5.2. The study's procedure including a table of all variables gathered at each of the three measurement points. PM = point of measurement; RQ = research question.

Procedure

Figure 5.2 depicts the study's procedure and all variables that were gathered at each of the three measurement points including their function concerning the four research questions (measures are described in more detail in section "Measures"). The study took place between spring 2013 and summer 2014 at four Swiss Gymnasien. In preparation for the study, all of the four participating teachers received a training (see section "Teacher Training"). Since lesson structure and curricula differ between schools in Switzerland, the four participating teachers individually started with the study as soon as their physics classes had to be taught introductory mechanics. Therefore, also the time gap between the teacher training and the implementation of the CogAct teaching unit varied between the teachers (from a few days to half a year).

About one week before each of the eight participating classes started with introductory Newtonian mechanics, the teachers received a link to an online-questionnaire eliciting demographic and motivational variables (see section "Measures") that had to be forwarded to the students. The students were instructed to fill in the questionnaire within one week. Immediately before each of the eight participating classes started with introductory Newtonian mechanics, the first author assessed the students' prior conceptual understanding regarding Newtonian mechanics. After the pretesting, the classes received 18 lessons (à 45 minutes) CogAct instruction or conventional instruction, respectively. During that time, the first author stayed in close contact with the teachers to provide ongoing support if needed and to monitor and increase implementation fidelity. The first author further made unannounced classroom visits (once in each classroom) protocolled systematically to get an idea of the CogAct teaching unit's implementation and of the teachers' instruction in the conventional classrooms.

Immediately after the 18 lessons, at the posttest, the students were invited to fill in a second online-questionnaire that again assessed the same motivational variables that were already assessed before instruction. Also conceptual understanding was measured a second time, while quantitative problem solving performance was registered for the first time.

Independent of condition, teachers had to take care of proper examinations and grading themselves, according to their school-specific requirements. The tests that were applied in the context of the study did not substitute regular exams. To also have a measure of school

achievement, the students' physics grades that reflected the students' performance in introductory Newtonian mechanics (i.e., the content covered during the 18 lessons) were recorded. Importantly, all teachers were requested to schedule the main regular exam as close to the study's posttest as possible to ensure comparable external learning conditions at posttest across classes.

Approximately three months after the completion of the 18 lessons, conceptual understanding and quantitative problem solving performance were elicited once again. At any time before or after the 18 lessons, the first author administered an intelligence test in each of the eight classes.

Teacher Training

The teachers were trained in a way suggested to increase implementation fidelity by existing research (e.g., Furtak et al., 2008; Greene, 2015; Harris et al., 2015). In a two-day training carried out by all of the authors including an in-service physics teacher, the teachers learnt about the theoretical ideas behind the CogAct teaching unit and were introduced to the CogAct teaching manual that explicates the whole teaching unit. The structure and usage of the manual and the attached additional worksheets and power-point slides were described (the unit is introduced in more detail in section "The CogAct Teaching Unit on Basic Newtonian Mechanics"). The manual that could be employed similar to a script provided clear guidance on how the CogAct teaching unit should be implemented. Every lesson could be taught only relying on the available information and materials. We provided the highly structured manual including all necessary teaching materials as a guideline for the teachers and to reduce cognitive demands during the implementation. At the same time, however, the teachers should understand what they implement and why and be able to adjust the teaching unit to their own teaching preferences while being in keeping with its theoretical ideas. This is why the training emphasized the communication of the ideas behind the CogAct methods and principles included in the unit and also incorporated a discussion of the teaching unit's implementation in the context of the study. In this discussion, the participating teachers could contribute to the interpretation of the manual regarding the study and solutions to several important questions were developed together. These questions included, for instance, what elements of the CogAct teaching unit can and cannot be omitted or how much leeway is necessary to adapt the teaching to the students' needs. The CogAct methods in the unit,

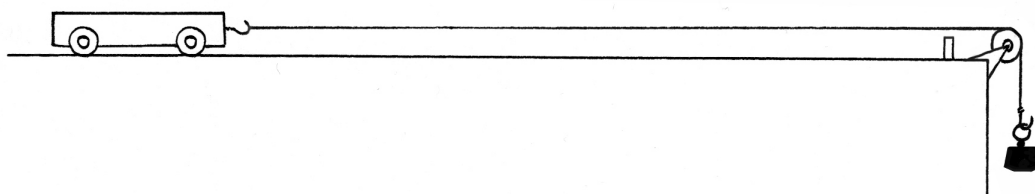
aimed at fostering conceptual understanding, can be described as *active ingredients* (Greene, 2015) that are expected to be crucial for the whole intervention's effectiveness. Although the teachers were free to modify the suggested methods, develop their own methods, or choose from several proposed methods as long as the idea behind the respective method (e.g., holistic mental model confrontation or metacognitive reflection) was retained, it was clearly communicated that these instructional elements are the *active ingredients* of the CogAct teaching unit and must not be omitted. Because the specific sequencing of the topics in the CogAct teaching unit was intended to promote the active incremental construction of knowledge, the teachers were further requested to stick to the given order. All teachers received a protocol documenting the results of the discussion.

The training also informed the teachers about the study's time schedule and associated obligations. We presented all the mechanics topics covered in the CogAct teaching unit that had to be taught in the conventional instruction (the control group) as well. In terms of the conventional instruction, the teachers were told that they should teach introductory Newtonian mechanics as always with the only restriction that all topics presented had to be covered within the study's time frame of 18 lessons (just as in the CogAct instruction), in an individual order and with individual prioritization.

The CogAct Teaching Unit on Basic Newtonian Mechanics

The CogAct teaching unit on basic Newtonian mechanics was created by teaching experts at the MINT-Learning Center ETH Zurich. CogAct instructional methods are applied to develop new conceptual knowledge and to overcome unfavorable prior knowledge during instruction (e.g., inventing, holistic mental model confrontation), but also to rework and elaborate content and monitor own learning processes at home (e.g., self-explanations, metacognitive questions). The learners are required to actively deal with the content to be learnt. Regarding the general structure of the program, explorative and explanative elements alternate. Thus, new topics are introduced with "unexplainable" phenomena that are intended to trigger questions that connect to the new topic. Correspondingly, the unit is organized in terms of questions that are stimulated to come up during the lessons and answered later on. In general, the sequence of the topics is chosen in a way that each topic follows naturally from the preceding topic to help the students build coherent and solid knowledge structures that make sense to them and therefore facilitate active knowledge construction. Working with

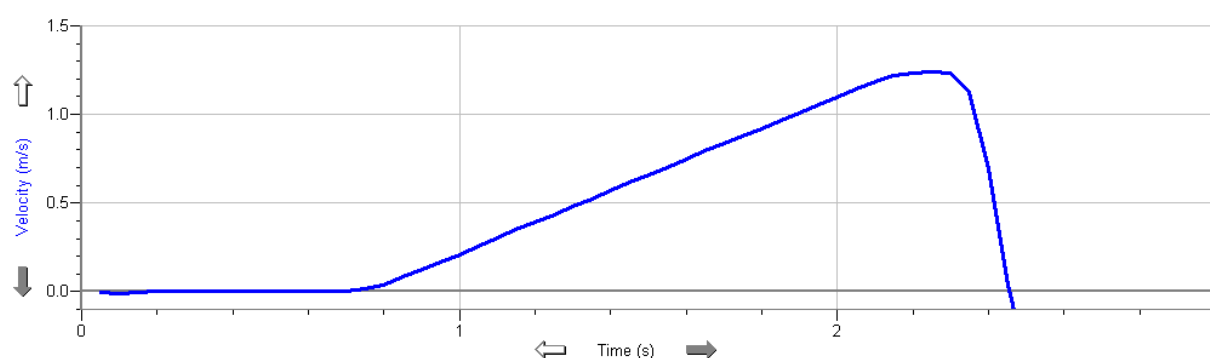
observable real-world applications that connect to existing knowledge and are intended to raise motivation, the unit implements CogAct methods to explicitly prompt students to think on their own and participate in discussions in order to develop conceptual knowledge.



The students had already learnt how a body moves without a net force acting upon it.

What do the students expect in terms of the movement of the wagon in the above arrangement?

Usually, the expectations of the students are diverse indicating that the concept of inertia has not been fully understood so far. By asking the students about their expectations, misunderstandings can be detected and addressed and existing beneficial knowledge is consolidated and activated in preparation for the new information to be acquired.



After collecting and discussing the students' expectations, the wagon's movement is plotted. How can this movement be described? The students are asked to invent a common index that describes the movement of the wagon. In this individual phase and the following discussion, the students are promoted to acquire the concept of acceleration themselves.

Figure 5.3. Condensed excerpt of a learning sequence on the topic acceleration from the CogAct teaching unit on basic Newtonian mechanics.

Importantly, the unit focusses on the observation and description of phenomena, while the formalization is to happen along the way. Accordingly, instructional time is spent predominantly on developing a conceptual understanding of mechanics contents. Starting from a conceptual understanding, students learn en passant how concepts translate into formalisms to be able to implement them quantitatively. Quantitative problem solving is intended to succeed primarily based on the understanding of the underlying concepts. Therefore, considerably less time is devoted to practicing quantitative problem solving.

The CogAct teaching unit consists of six parts encompassing altogether 16 lessons (plus two extra lessons as buffer time). The six parts cover “inertia and motion”, “mass and weight”, “force and acceleration”, “balance of forces”, “reciprocal action”, and, finally, “Newton’s axioms”. Each lesson is introduced according to the principles described above and provides several possibilities for in-depth studies to allow adaption to classroom needs. To exemplify the general idea realized in the CogAct teaching unit on introductory Newtonian mechanics, Figure 5.3 provides a condensed excerpt of a learning sequence on the topic acceleration.

Measures

All variables measured in this study are listed in Figure 5.2. The students’ age, gender, and specialization at school on non-STEM fields vs. STEM fields were assessed by means of the online-questionnaire administered at pretest. School achievement was gathered in the form of the students’ physics grades reflecting performance in introductory Newtonian mechanics (i.e., the content covered during the 18 lessons). In Switzerland grades range from 1 to 6 with smaller numbers indicating lower performance. With grades lower than 4 students fail. All other variables and their implementation are described in more detail.

Motivational variables. The scales that were included in the online-questionnaire at pre- and posttest to measure students’ interest in physics, physics self-concept, use of efficient learning strategies in physics, physics learning amotivation, and physics anxiety are presented in the following (see Appendix A).

Interest in physics. The scale to measure interest in physics was adopted from the international student-survey of PISA 2006 (Frey et al., 2009). It consisted of four items with

four-point Likert scales spanning from 0 “completely disagree” to 3 “completely agree” (Cronbach’s $\alpha = .91$; sample item: “These days I like dealing with physics problems.”).

Self-concept in physics. The students’ self-concept in physics was elicited adapting four items of the “DISK-Gitter mit SKSLF-8” (Rost, Sparfeldt, & Schilling, 2007), a published test in German language targeting school subject specific self-concept. Students could choose between six answer alternatives spanning from 0 “not true for me at all” to 5 “exactly true for me” (Cronbach’s $\alpha = .91$; sample item: “These days I feel that I can solve problems in physics easily.”).

Learning strategies in physics. To capture the students’ learning strategies, always three items of the German PISA 2006 scales (Frey et al., 2009) targeting elaboration (sample item: “These days in physics lessons I visualize the content with examples.”) and organizational processes (sample item: “In these days’ physics instruction I try to recognize interrelations.”) were included. Control strategies (sample item: “These days I ask myself questions to make sure that I have understood the material covered in the physics lesson.”) were additionally assessed by combining two items from the German PISA 2000 control strategies scale (Kunter et al., 2002) and one item from the metacognitive self-regulation subscale of the “Motivated Strategies for Learning Questionnaire” (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1991; validated German version by Nenniger & Nyberg, 1992). Participants could choose from four answer alternatives (0 “almost never” to 3 “almost always”). Because the three scales were highly correlated and could not be distinguished statistically, they were merged to form one joint learning strategies scale (Cronbach’s $\alpha = .79$).

Learning amotivation in physics. Obtained from the German PISA 2006 student-survey (Frey et al., 2009) and then adjusted, three items measured learning amotivation in physics with four answer alternatives ranging from 0 “almost never” to 3 “almost always” (Cronbach’s $\alpha = .79$; sample item: “In these days’ physics lessons I don’t want to participate.”).

Physics anxiety. To measure physics anxiety, a version of the Abbreviated Math Anxiety Scale (Hopko, Mahadevan, Bare, & Hunt, 2003) adapted to physics instruction was administered. The students had to rate how anxious they would feel during the event described in each of the nine items, using a five-point Likert scale from 0 “low anxiety” to 4

“high anxiety” (Cronbach’s $\alpha = .80$; sample item: “Thinking about an upcoming physics test one day before.”).

Conceptual understanding. Conceptual understanding in introductory Newtonian mechanics was measured using a validated multiple choice test that was augmented by additional items at the post- and follow-up test (see Appendix B).

The basic Mechanics Concept Test. Prior conceptual understanding was assessed with the basic Mechanics Concept Test (bMCT), a short multiple choice test. It captures conceptual understanding of basic Newtonian mechanics. Only if no wrong answer alternative and all correct answer alternatives are checked, an item is scored one point. The test was administered with a maximal working time of 35 minutes ensuring that there was enough time for all students to work on and think through all of the problems. Figure 5.4 illustrates two sample test items. Hofer, Schumacher, and Rubin (2015) describe the test’s development and evaluation. The bMCT satisfies the Rasch model allowing the simple use of sum scores to adequately measure the students’ conceptual understanding. Since only the 11-items version of the bMCT can be validly applied both as pre- and posttest to measure development on one underlying dimension (Hofer et al., 2015), the 11-items version was used in this study (maximum score = 11). The problem contexts implemented in the bMCT were explicitly not discussed during the 18 lessons of introductory Newtonian mechanics instruction, neither in the CogAct nor in the conventional classes. Therefore, all items of the bMCT required the students to apply their conceptual understanding to new situations. The bMCT could hence be considered to assess deep conceptual knowledge that can be employed flexibly and transferred to superficially new problem situations (i.e., a conceptual transfer measure).

8. A person is standing in a resting boat and tosses a big stone into the water behind the boat. Which of the following statements are true?



- ☐ The boat moves in the direction the stone was thrown.
 - ☐ The stone displaces water and this is why the boat moves just slightly back and forth.
 - ☐ If you let an inflated balloon whizz through the air, principally the same happens.
 - ☐ The boat moves contrary to the direction the stone was thrown.
11. The following three balls are moving on a horizontal plane:
- Ball A is moving around a bend with a velocity of $1\frac{m}{s}$.
 - Ball B starts with a velocity of $6\frac{m}{s}$ and then becomes slower and slower.
 - Ball C is moving faster and faster.

Which of the following statements are true?

- ☐ A horizontal force is acting on ball A.
- ☐ A horizontal force is acting on ball B.
- ☐ A horizontal force is acting on ball C.

Figure 5.4. Two sample items of the basic Mechanics Concept Test (bMCT) translated into English. For item 8, the last two answer alternatives are correct and for item 11, all three answer alternatives are correct.

The bMCT plus. In the post and follow-up testing, the students' conceptual understanding was assessed with the bMCT augmented by six additional multiple choice items that resembled the original bMCT items. Hence, 17 points could be achieved at the maximum in the resulting bMCT plus. The six new items required the students to transfer their knowledge to another knowledge domain (e.g., transfer action-reaction principle from mechanics to magnetism) or to combine what they had learnt in the context of complex problem situations with several forces operating (e.g., elevator ride or tug-of-war). These items could be considered impossible to solve correctly without instruction. The bMCT plus that included a total of 17 items still had to be completed in maximally 35 minutes. This time limit, however, turned out to suffice to finish the test without having to hurry.

Quantitative problem solving. To examine quantitative problem solving performance, the students completed an additional test at the post and follow-up assessment (see Appendix C). The test included five quantitative problems targeting Newton's three axioms that required students to read graphs, apply formulae, and calculate. It closely resembled standard physics examinations. For at least ten minutes, the students worked on the quantitative problem solving test that immediately followed the assessment of conceptual understanding (i.e., the bMCT plus). Both assessments had to be completed in 45 minutes (i.e., one lesson). The test was scored according to a systematic scoring system coordinated with physics teaching experts (maximum score = 11.25). Two independent raters coded 32 tests à five quantitative physics problems according to the scoring system. The intra-class correlation coefficient confirmed high inter-rater agreement ($ICC = .91$). Hence, one of the two raters coded all of the tests according to the scoring system.

Intelligence. The students' intelligence was estimated by means of the well-established set II score of Raven's Advanced Progressive Matrices (maximum score = 36; Raven, Raven, & Court, 1992). Set I was used as training set and time on set II was limited to 40 minutes. The test was administered following the instructions described in the test's manual.

Data Analysis

Figure 5.2 also specifies the function of each variable regarding the four research questions. When introducing the statistical methods applied to analyze the variables corresponding to the four research questions, we hence point to Figure 5.2 for orientation

purposes. Mplus Version 7.11 (Muthén & Muthén, 2012) was used for all analyses. We conducted robust maximum likelihood estimation to potentially correct fit statistics and all parameter estimates' standard errors for leptokurtic or platykurtic data. If not otherwise specified, missing values were estimated using full information maximum likelihood (FIML).

Research question 1: General effectiveness. To investigate the general effectiveness of the CogAct instruction in terms of immediate (post) and long-term (follow-up) effects, two regression models were constructed. The conceptual understanding and quantitative problem solving post scores served as dependent variables in Model Post. They were regressed on condition (0 = conventional instruction, 1 = CogAct instruction) and the five control variables age (in years), gender (0 = female, 1 = male), specialization (0 = non-STEM, 1 = STEM), intelligence, and prior conceptual understanding. The five control variables were included to control for variations on the individual student level that could not be attributed to the intervention but had to be considered as additional predictors of learning due to the quasi-experimental setting. The second regression model, Model Follow-up, exactly resembled Model Post with the only exception that the conceptual understanding and quantitative problem solving follow-up scores were used as dependent variables. Consequently, two regressions were estimated at posttest and two regressions were estimated at follow-up test. With Y representing the respective dependent variable, the four regressions hence read as follows: $Y = \beta_0 + \beta_1 * \text{condition} + \beta_2 * \text{age} + \beta_3 * \text{gender} + \beta_4 * \text{specialization} + \beta_5 * \text{intelligence} + \beta_6 * \text{prior conceptual understanding} + \varepsilon$. Significant positive regression coefficients of condition (β_1) indicated learning advantages of CogAct instruction over conventional instruction.

The p -values resulting from the significance tests of the regression coefficients may be distorted since they are based on the assumption of normally distributed parameters. To get more stable p -values, we performed log-likelihood tests that, in general, compare less restrictive models to more restrictive but nested models (for detailed information on the test, see UCLA: Statistical Consulting Group, 2014). Hence, a chi-square distributed test statistic resulting in a p -value ($LL\ p$) is calculated for the models that are compared, using log-likelihoods, scaling correction factors, and the numbers of free parameters. Significant discrepancies in model-fit indicate that the more restrictive model fits the data significantly worse than the less restrictive model. When the significance of regression coefficients is to be evaluated, the respective regression coefficient is set to zero in the restrictive model and

estimated freely in the unrestrictive model. A significant *LL* *p*-value then suggests that the regression coefficient significantly contributes to the regression model and should not be set to zero. We inspected *LL* *p*-values to determine the significance of regression coefficients in all regression models analyzed in this study.

Research question 2: Accessing procedures via concepts. Scaffolds were included randomly into half of all quantitative problem solving tests within each class at both post and follow-up testing. These scaffolds simply prompted the students to think about and write down the physics terms and principles that have to be considered before they start calculating. We hypothesized that, if not all CogAct students, at least scaffolded CogAct students should outperform conventional learners in terms of their quantitative problem solving. The effect of scaffolding was investigated using regression models again. We examined the interaction between the two independent variables condition (conventional = 0, CogAct = 1) and posttest scaffold (no scaffold = 0, scaffold = 1) in terms of immediate effects on posttest quantitative problem solving and in terms of delayed effects on follow-up quantitative problem solving, adjusting for the five control variables. An analogue regression model was run for the follow-up scaffold and its effects on follow-up quantitative problem solving. With *Y* representing posttest or follow-up quantitative problem solving, the regressions read as follows: $Y = \beta_0 + \beta_1 * \text{condition} + \beta_2 * (\text{posttest/follow-up}) \text{ scaffold} + \beta_3 * \text{condition} \times (\text{posttest/follow-up}) \text{ scaffold} + \beta_4 * \text{age} + \beta_5 * \text{gender} + \beta_6 * \text{specialization} + \beta_7 * \text{intelligence} + \beta_8 * \text{prior conceptual understanding} + \varepsilon$. Significant positive regression coefficients of condition \times (posttest/follow-up) scaffold (β_3) indicated that CogAct students profited more from the scaffolding than conventional students.

Importantly, while the scaffold was assumed to specifically help CogAct learners to access problem solving procedures by first activating the respective concepts, we did not expect the scaffolding to differently influence the conceptual understanding of CogAct and conventional learners. At both post and follow-up testing, quantitative problem solving was examined after the assessment of conceptual understanding. Hence, immediate effects of scaffolding on conceptual understanding could not be investigated. Nevertheless, delayed effects of posttest scaffolding could potentially be observed on conceptual understanding at follow-up. Therefore, an analogue regression model additionally checked for potential effects of posttest scaffolding on follow-up conceptual understanding.

Research question 3: Impact on motivation. Does CogAct instruction increase the students' interest in physics, physics self-concept, and use of efficient learning strategies as well as decrease the students' learning amotivation and physics anxiety as compared to conventional instruction? To investigate the third research question, five regression analyses were performed. The five motivational variables as measured at posttest served as dependent variables. Each single motivational variable was regressed on condition and the same variable measured at pretest to control for prior differences: $Y = \beta_0 + \beta_1 * \text{condition} + \beta_2 * \text{motivation pretest} + \varepsilon$. Significant positive regression coefficients of condition (β_1) indicated motivational advantages of CogAct instruction over conventional instruction.

Research question 4: Impact on physics underachievement. Can physics instruction based on the CogAct teaching unit tackle underachievement? The analytical strategy chosen to answer this research question is presented in the following sections.

Comparison of profile structure and probability to be an underachiever. To be able to compare the profile structure and the probability to be an underachiever between CogAct and conventional instruction, multiple group latent profile analyses (LPAs) were conducted. In general, to define underachievers, the z-standardized intelligence and the z-standardized physics grades were used as profile indicator variables, while the multiple groups were determined by the students' gender as moderating variable. For these analyses, we applied listwise exclusion of missing values because the student profiles could only be estimated validly with both indicator variables at hand. A systematic co-occurrence of high intelligence and low physics grades indicated a profile of physics underachievers. Detailed information on the examination of physics underachievement by means of LPA is provided by Hofer and Stern (2015). According to these authors, a five-profiles-solution with one of the profiles, physics high achievers, showing measurement invariance across gender was the best fitting model. One of the resulting five student profiles corresponded to a physics underachievers profile. This profile was detected only among female students. To examine whether underachievers existed in both conditions, we compared the profile structure between CogAct and conventional instruction. Hence, the five-profiles-solution was realized within the sample of students having received CogAct instruction and within the sample of students having received conventional instruction. The fit of a model that allowed the profiles to be estimated independently for CogAct and conventional students was compared with the fit of a model that constrained the profiles to be equal, using log-likelihood tests. No significant

discrepancies in model-fit indicated that the model with equated profiles did not fit the data significantly worse than the unrestricted model. Such a finding would suggest that the same student profiles, including the female physics underachievers, were present in both conditions. On the contrary, significant discrepancies in model-fit would suggest a different profile structure depending on the instructional condition. Such an outcome would require a detailed inspection of the resulting student profiles within each condition to find out whether an underachievers profile existed in the CogAct instruction condition or not.

In case of no significant differences in the profile structure between conventional and CogAct instruction, the probability to be an underachiever was compared by realizing the five-profiles-solution (with one profile constrained to be equal between female and male students) in the total student sample and subsequently predicting the students' estimated profile membership probabilities by condition (conventional = 0, CogAct = 1). A significant regression coefficient of condition indicated differences in profile membership probabilities between CogAct instruction and conventional instruction. A significant negative regression coefficient estimated for the latent profile category of the underachieving girls indicated a lower probability to be in the female underachievers profile for students having received CogAct instruction than for students having received conventional instruction. The inclusion of the predictor condition should not affect the profile estimation. Hence, the profiles were fixed according to the manual 3-step approach as described by Asparouhov and Muthén (2012) before including condition as predictor variable. While deterministically categorizing students based on a most-likely-latent-profile-membership variable and performing a regression analysis afterwards lead to results afflicted with disregarded categorization errors, the 3-step approach considers the probabilistic nature of profile membership.

Analyses on study performance measures and motivation. Yet, potential effects of the CogAct intervention might not be reflected in the students' physics grades, resulting in no significant differences between conventional and CogAct instruction regarding the existence and relative frequency of physics underachievers. In that case, we additionally planned to investigate the underachieving girls' performance in terms of their conceptual understanding and quantitative problem solving as well as their physics motivation (i.e., interest in physics, physics self-concept, learning strategies in physics, learning amotivation in physics, and physics anxiety). Hence, we planned to estimate the mean conceptual understanding and quantitative problem solving posttest and follow-up scores as well as the mean manifestations

on the motivational variables within the student profiles and separately for each condition. To this end, the profiles were fixed again before estimating the profile-specific means and variances for each condition applying the manual 3-step approach (Asparouhov & Muthén, 2012). The estimated means and variances could then be compared for underachieving female students having received CogAct vs. conventional instruction. To test whether the estimated means and variances (i.e., $df = 2$) significantly differed between CogAct and conventional instruction, the chi-square value (χ^2) of the Wald test of parameter constraints was inspected.

Results

As a preliminary remark, communication with the teachers and classroom visits indicated that the teachers managed to handle the CogAct instruction based on the CogAct teaching unit quite well. Yet, we also registered some initial difficulties with attuning personal preferences to the unit's standards and a decrease in authenticity and fluency in the teaching process as compared to instruction as always. Still, overall, the CogAct instruction including CogAct methods was implemented as intended suggesting rather high implementation fidelity. After the presentation of the descriptive statistics, the results are described according to the four research questions.

Descriptive Statistics

Table 5.1 summarizes the descriptive statistics of the major continuous study variables organized by instructional condition. Regarding the dichotomous study variables gender and specialization, 52% of all conventional students and 55% of all CogAct students were females and 65% of all conventional students and 70% of all CogAct students specialized in a non-STEM subject.

Table 5.1

Condition-Specific Means, Standard Deviations, and Scales of Major Continuous Study Variables

Variables	Instructional Condition				Scale
	CogAct		Conventional		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Intelligence	27.36	4.56	27.41	4.05	0-36
Pre					
Age	16.00	0.81	15.92	1.10	
Interest	1.66	0.60	1.54	0.64	0-3
Self-concept	2.52	1.07	2.52	1.07	0-5
Learning strategies	1.64	0.51	1.60	0.50	0-3
Learning amotivation	0.85	0.55	0.78	0.57	0-3
Anxiety	0.82	0.63	0.90	0.54	0-4
Prior conceptual understanding	2.95	1.57	2.77	1.59	0-11
Post					
Interest	1.51	0.89	1.65	0.76	0-3
Self-concept	2.53	1.35	2.71	1.20	0-5
Learning strategies	1.67	0.50	1.61	0.60	0-3
Learning amotivation	1.21	0.81	0.95	0.73	0-3
Anxiety	0.74	0.63	0.84	0.59	0-4
Conceptual understanding	6.57	2.96	5.57	2.23	0-17
Quantitative problem solving	4.63	3.20	3.94	3.07	0-11.25
Physics grades	4.55	0.60	4.62	0.67	1-6
Follow-up					
Conceptual understanding	5.62	2.99	5.09	2.19	0-17
Quantitative problem solving	3.10	2.96	3.73	2.69	0-11.25

Research Question 1: General Effectiveness

To investigate the general effectiveness of the CogAct instruction in terms of immediate (post) and long-term (follow-up) effects on the two dependent variables conceptual understanding and quantitative problem solving, two regression models were constructed: Model Post and Model Follow-up. Table 5.2 presents the results of the analyses based on the two models. Importantly, students in the CogAct and conventional condition did not differ in terms of the control variables' means or proportions, respectively (all $ps \geq .45$).

Table 5.2

Parameter Estimates Based on the Regression Model for Post Data and the Regression Model for Follow-Up Data

Variables	Model Post			Model Follow-up		
	<i>b</i>	<i>SE</i>	<i>LL p</i>	<i>b</i>	<i>SE</i>	<i>LL p</i>
DV = Conceptual understanding						
Condition (0 = conventional, 1 = CogAct)	1.03	0.36	<.01	0.68	0.36	<.05
Control variables						
Age	0.09	0.22	.67	0.11	0.21	.59
Gender (0 = female, 1 = male)	1.06	0.37	<.01	1.24	0.37	<.01
Specialization (0 = non-STEM, 1 = STEM)	0.52	0.42	.21	0.46	0.45	.28
Intelligence	0.07	0.04	.10	0.05	0.03	.18
Prior conceptual understanding	0.57	0.14	<.001	0.70	0.14	<.001
DV = Quantitative problem solving						
Condition	0.87	0.46	<.05	-0.50	0.45	.28
Control variables						
Age	-0.15	0.26	.56	-0.09	0.25	.72
Gender	0.33	0.48	.50	0.53	0.45	.25
Specialization	0.31	0.53	.56	0.88	0.54	.09
Intelligence	0.21	0.04	<.001	0.05	0.04	.21
Prior conceptual understanding	0.36	0.17	<.05	0.36	0.16	<.05

Note. DV = dependent variable; *LL p* = *p*-values that resulted from the log-likelihood tests.

In Model Post, being in the CogAct condition had a significant positive effect both on conceptual understanding ($\beta = 0.19$, $SE = 0.06$) and quantitative problem solving ($\beta = 0.14$, $SE = 0.07$). This implied an advantage of 1.03 points (95% CI [0.32, 1.73]) in the conceptual understanding test and an advantage of 0.87 points (95% CI [-0.02, 1.77]) in the quantitative problem solving test for CogAct students. In Model Follow-up, students only profited significantly from CogAct instruction in terms of conceptual understanding ($\beta = 0.13$, $SE = 0.07$), indicating an advantage of 0.68 points (95% CI [-0.02, 1.39]) in the conceptual

understanding test. Importantly, these effects were present after controlling for the five individual student variables. Figure 5.5 depicts the estimated means of post and follow-up conceptual understanding (5.5a) and post and follow-up quantitative problem solving (5.5b) as a function of condition, with all control variables set at their means.

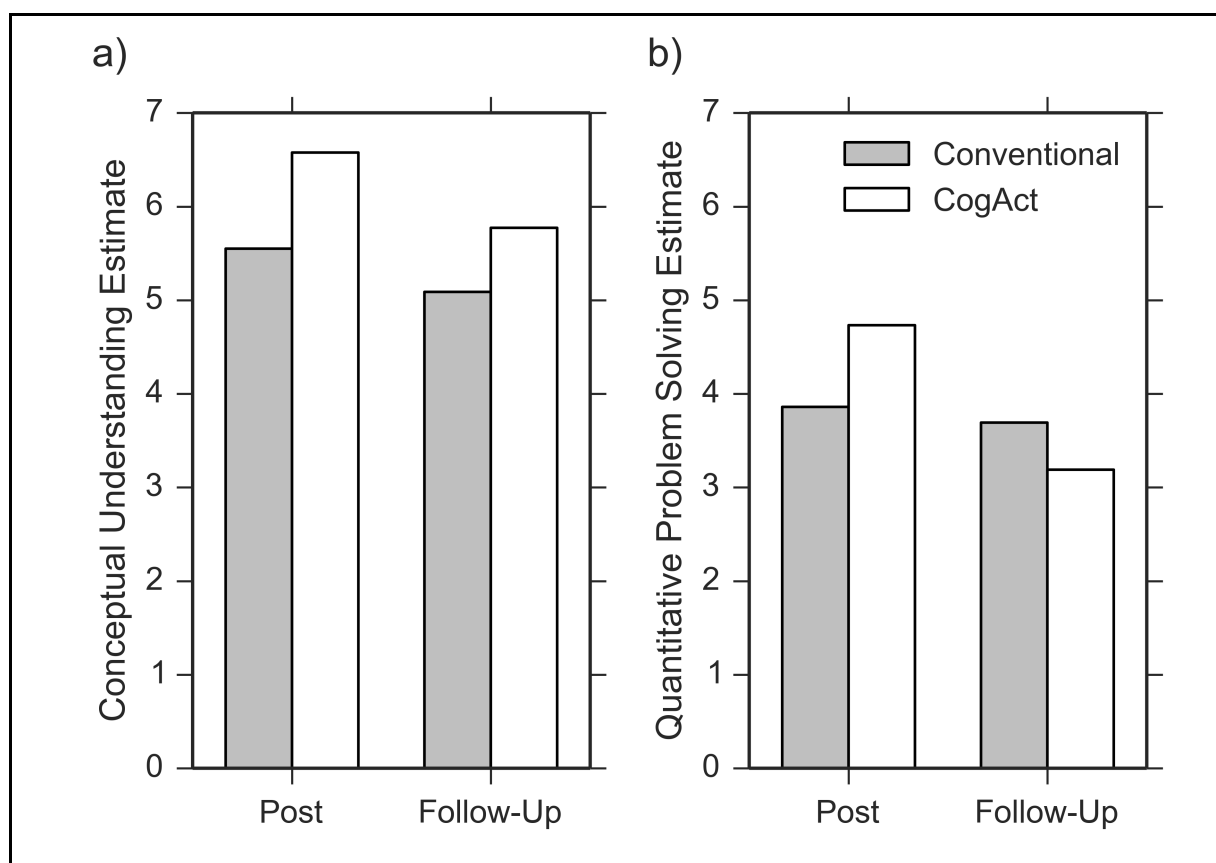


Figure 5.5. Conceptual understanding (a) and quantitative problem solving estimates (b) for post- and follow-up tests as a function of condition based on the regressions listed in Table 5.2. The intercepts are $b_{0,\text{post}} = -0.01$ and $b_{0,\text{follow-up}} = -0.63$ for conceptual understanding and $b_{0,\text{post}} = -0.60$ and $b_{0,\text{follow-up}} = 2.14$ for quantitative problem solving. Control variables are set at their means.

Research Question 2: Accessing Procedures via Concepts

Using regression models, we examined the interaction between the two independent variables condition (conventional = 0, CogAct = 1) and posttest scaffold (no scaffold = 0, scaffold = 1) in terms of immediate effects on posttest quantitative problem solving and in

terms of delayed effects on follow-up quantitative problem solving, accounting for the five control variables. An analogue model was run for the follow-up scaffold and its effects on follow-up quantitative problem solving. Because all these analyses involved subgroups leading to smaller group sizes, the student sample was augmented by pilot data. The pilot sample included data from two physics classrooms instructed as always and four classrooms instructed according to the CogAct teaching unit. The main sample and the pilot sample were highly comparable, however, they differed in the important aspect that in the pilot study each teacher taught either conventionally or according to the CogAct teaching unit. Hence, the teacher variable was not controlled in the pilot study. Yet, balancing this limitation against the problem of small group sizes, it was considered negligible in the context of these particular analyses due to the randomization of the scaffolding within each classroom. The pilot sample included $N = 113$ (67 girls) Swiss students with a mean age of $M = 15.79$ years ($SD = 0.95$ years), resulting in a total sample of $N = 285$ students. Importantly, whenever scaffolds were provided but all respective fields in the test were left blank (no ideas written down), we assumed that the student had ignored the scaffolding and excluded the student from the analyses. The finally resulting group sizes are listed in Table 5.3.

Table 5.3

Group Sizes for Analyses on the Effects of the Interaction between Condition and Scaffold (Condition \times Scaffold) on Quantitative Problem Solving

Instructional condition	Post		Follow-up	
	No scaffold	Scaffold	No scaffold	Scaffold
Conventional	48	41	54	39
CogAct	58	64	60	48

The interaction between condition and scaffold that indicated whether CogAct students and conventional students profited differently from the scaffolding, was only significant regarding posttest scaffolding and follow-up quantitative problem solving ($b = 1.98$, $SE = 0.98$, 95% CI [0.06, 3.89], $LL p < .05$; $\beta = 0.29$, $SE = 0.15$). Both main effects (condition and posttest scaffold) were not significant (all $LL ps \geq .10$). Hence, CogAct students profited more from posttest scaffolding in terms of its effect on follow-up quantitative problem solving than conventional students. The follow-up quantitative problem solving estimates

resulting from this regression model are visualized in Figure 5.6 as a function of condition and posttest scaffold. Neither did the follow-up scaffold affect follow-up quantitative problem solving nor did the posttest scaffold affect posttest quantitative problem solving. The posttest scaffold also had no influence on the follow-up conceptual understanding (for all main effects of scaffolding and the interaction effects between condition and scaffold, all LL $ps \geq .31$).

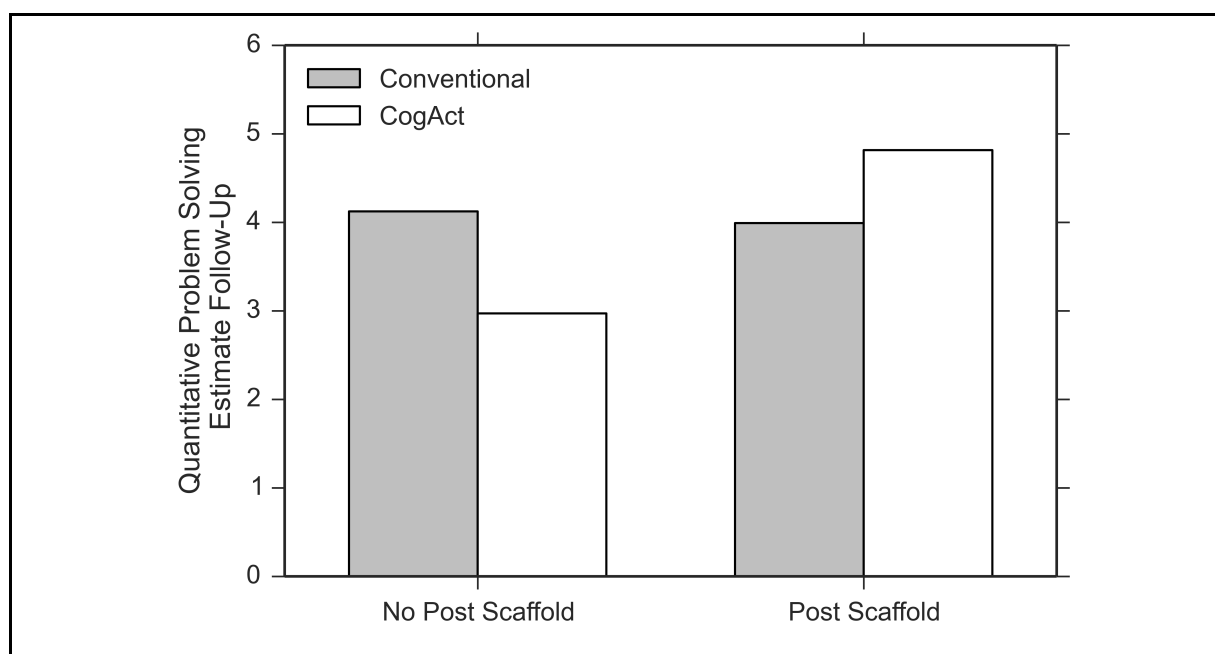


Figure 5.6. Follow-up quantitative problem solving estimates as a function of condition and posttest scaffold. The follow-up quantitative problem solving estimates Y are calculated based on the regression equation $Y = 2.55 + (-1.15) * \text{condition} + (-0.13) * \text{posttest scaffold} + 1.98 * \text{condition} \times \text{posttest scaffold} + (-0.13) * \text{age} + 1.02 * \text{gender} + 0.07 * \text{specialization} + 0.08 * \text{intelligence} + 0.38 * \text{prior conceptual understanding}$. Control variables are set at their means.

Research Question 3: Impact on Motivation

Regarding the CogAct instruction's impact on the motivational variables interest in physics, physics self-concept, learning strategies in physics, learning amotivation in physics, and physics anxiety, the regression coefficient of condition was not significant for any of the motivational variables when controlling for the pretest scores on the respective motivational

variables (all $LL\ ps \geq .12$). Hence, no motivational advantages of CogAct instruction over conventional instruction were found.

Research Question 4: Impact on Physics Underachievement

Comparison of profile structure and probability to be an underachiever. To investigate the influence of CogAct instruction on physics underachievement, a multiple group LPA was conducted to replicate the five-profiles-solution described by Hofer and Stern (2015) with the z-standardized intelligence and the z-standardized physics grades as profile indicator variables and the students' gender as moderating variable. According to information criteria (Akaike, Bayesian, and sample-size adjusted Bayesian), the a priori favored five-profiles-solution fitted the data better than solutions with one to 14 profiles. The log-likelihood test indicated that the five profiles did not differ significantly for students having received CogAct instruction and students having received conventional instruction ($LL\ p = .31$). The information criteria suggested the restrictive (CogAct = conventional) solution, too.

Using the total student sample, we could replicate the five student profiles identified by Hofer and Stern (2015): a high achievers profile ($N = 63$; 20 girls) being invariant for girls and boys (very high intellectual potential, $M = 0.64$, and very high physics grades, $M = 0.72$), girls' underachievers ($N = 25$; high intellectual potential, $M = 0.44$, and very low physics grades, $M = -1.23$), girls' over-to-normal achievers ($N = 31$; rather low intellectual potential, $M = -0.99$, and average physics grades, $M = -0.13$), boys' under-to-normal achievers ($N = 14$; average intellectual potential, $M = -0.08$, and rather low physics grades, $M = -0.85$), and boys' overachievers ($N = 16$; very low intellectual potential, $M = -1.28$, and average physics grades, $M = 0.07$). Standard deviations, by default, were constrained to be equal across the profiles to simplify the model. Hence, the overall standard deviation was $SD = 0.65$ for intellectual potential and $SD = 0.67$ for physics grades.

Importantly, the most likely latent profile membership of each student indicated that there was no accumulation of particular student profiles (e.g., underachievers, high achievers) within specific conventional or CogAct classrooms. Consequently, the student profiles were not classroom-dependent. Furthermore, it has to be noted that the sample sizes given for each of the student profiles for orientation purposes were also based on the most likely latent profile membership patterns. In the following analyses, however, profile membership

probabilities were used for calculations. We did not analyze distinct subgroups. Although we speak of the student profiles of the high achievers or underachievers as groups of students, the probabilistic nature of all of these profiles should still be kept in mind.

To compare the probability to be an underachiever between CogAct and conventional instruction, the students' estimated profile membership probabilities were regressed on condition (conventional = 0, CogAct = 1). However, condition turned out to be no significant predictor of profile membership probabilities ($LL\ p = .30$). In particular, also the probability to be in the female underachievers profile could not be predicted by condition (the smallest p -value, $p = .08$, resulted for parameterization using the girls' high achievers as reference). Consequently, the analyses suggested that CogAct instruction had no beneficial influence on physics underachievement defined by intelligence and physics grades. This outcome allowed two possible conclusions: First, underachievers did not profit from CogAct instruction. Second, potential effects of the CogAct intervention might not be reflected in the students' physics grades. The following analyses aimed at clarifying this issue.

Analyses on study performance measures and motivation. Although the underachieving female students thus did not profit from CogAct instruction in terms of their school achievement (i.e., physics grades), we additionally investigated these girls' performance in terms of their conceptual understanding and quantitative problem solving as well as their physics motivation. CogAct instruction had no general beneficial effects on the students' motivational background in the overall sample. Nevertheless, positive effects could become apparent when CogAct instruction was compared with conventional instruction within the group of the female underachievers. We hence looked at the study performance and motivational measures within the student profiles as a function of instructional condition. For a more differentiated picture, female and male high achievers were considered separately instead of regarding only one joint high achievers profile. While the study performance measures are dealt with first, the motivational variables are addressed afterwards.

Performance. Figure 5.7 shows the means on posttest (5.7a) and follow-up conceptual understanding (5.7b) as well as on posttest quantitative problem solving (5.7c), estimated within the student profiles as a function of condition. Follow-up quantitative problem solving was omitted from this analysis due to the significant interaction between condition and posttest scaffolding. Hence, it would have been appropriate to additionally consider whether a scaffold was provided or not resulting in insufficient profile- and condition-specific sample

sizes. The estimated means for male under-to-normal achievers and overachievers are not represented in Figure 5.7 due to the small number of students in our sample having a high probability to be in one of these profiles and the resulting uncertainty in the estimation of condition-specific means. The 95% confidence intervals of the mean scores illustrate the coherence in the profile- and condition-specific estimates. We used the confidence intervals together with the overall patterns emerging from the graphs to describe the influence of CogAct instruction on the student profiles. In addition, we inspected the significance of the Wald test of parameter constraints comparing profile-specific means and variances between CogAct and conventional instruction, if relevant. Figure 5.7 shows that the female underachievers who had received CogAct instruction indeed managed to slightly gain on their higher achieving peers in terms of posttest conceptual understanding (5.7a) and even managed to catch up with the high achieving students at the follow-up test (5.7b). The conceptual understanding of the underachieving girls who had received conventional instruction, however, slightly lagged behind the performance of the other students at both measurement points and was significantly lower than the performance of the underachieving girls who had received CogAct instruction at follow-up ($\chi^2 = 30.46, p < .0001$). The female underachievers' posttest quantitative problem solving performance was not affected by the CogAct instruction.

Although we had no specific hypotheses concerning student profiles other than the female physics underachievers, the profile of the female high achievers stuck out. The analyses suggested considerable benefits of CogAct instruction for high achieving female students. The high achieving girls particularly profited from the CogAct instruction on all of the three performance measures and even caught up with their high achieving male counterparts (see Figure 5.7). The CogAct female high achievers significantly outperformed conventional female high achievers on posttest conceptual understanding ($\chi^2 = 12.90, p < .01$), on follow-up conceptual understanding ($\chi^2 = 7.15, p < .05$), and on posttest quantitative problem solving ($\chi^2 = 14.49, p < .001$). At the posttest, the high achieving females were even significantly better in quantitative problem solving than the high achieving males ($\chi^2 = 24.29, p < .001$). Importantly, at the same time, the high achieving boys were not handicapped by the CogAct instruction. On the contrary, CogAct instruction also significantly boosted the male high achievers' conceptual understanding at posttest as compared to conventional instruction (see Figure 5.7a; $\chi^2 = 11.01, p < .01$). However, the high achieving boys, in general, reached high scores on all performance measures independent of condition.

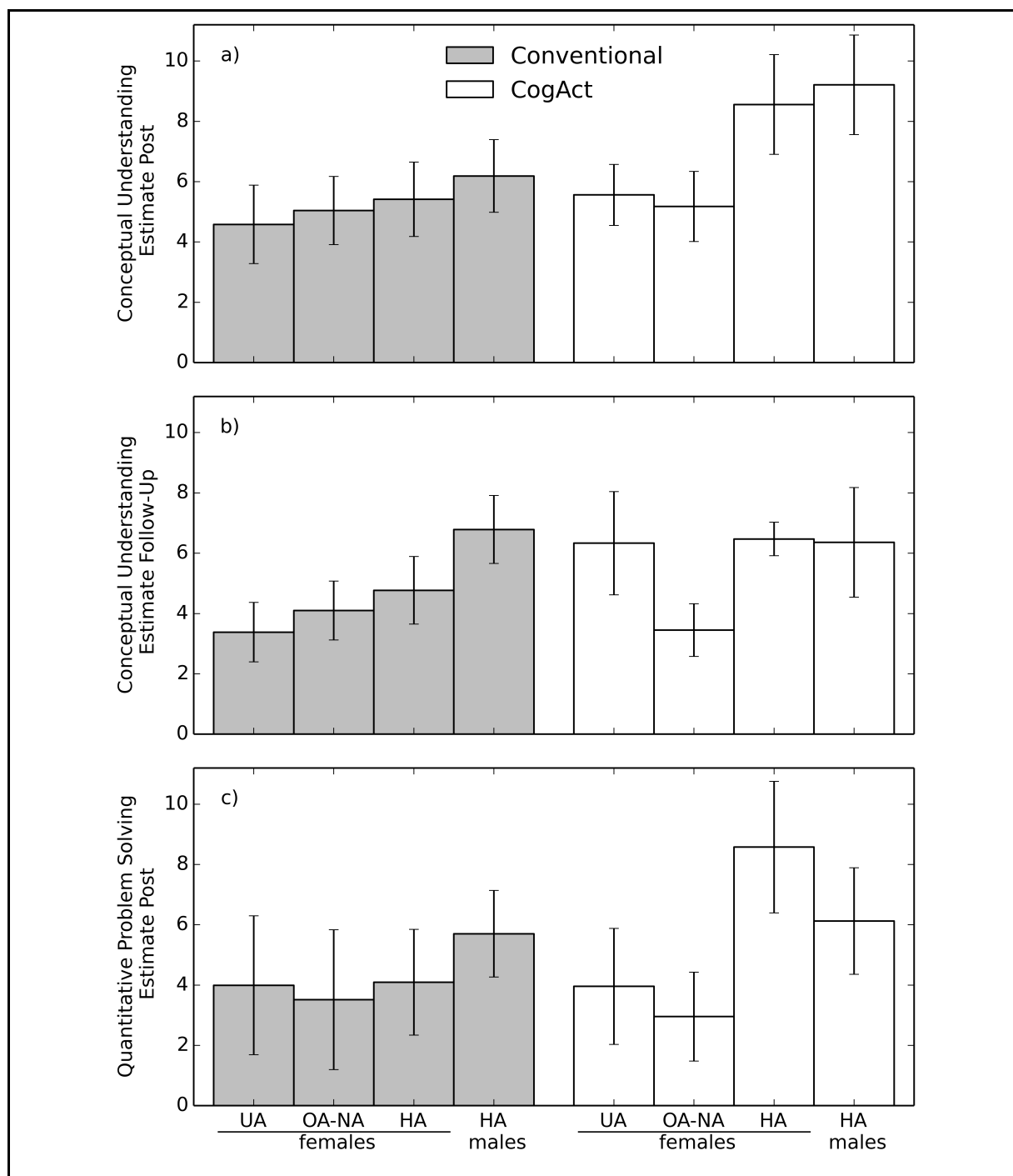


Figure 5.7. Posttest (a) and follow-up conceptual understanding (b) estimates and posttest quantitative problem solving estimate (c) as a function of student profile and condition. UA = underachievers, OA-NA = over-to-normal achievers, HA = high achievers. Error bars represent the 95% confidence intervals.

Motivation. The specific performance advantages of CogAct instruction for underachieving as well as high achieving girls were partially reflected in differences between CogAct and conventional instruction regarding some of the motivational variables. Hence,

underachieving female students having received CogAct instruction indicated using efficient learning strategies significantly more often ($M = 1.19$, $SD = 0.45$) than underachieving female students having received conventional instruction ($M = 0.48$, $SD = 0.19$; $\chi^2 = 11.36$, $p < .01$). They moreover showed a significantly lower learning amotivation ($M = 2.02$, $SD = 0.38$) than the conventional underachieving girls ($M = 2.42$, $SD = 0.15$; $\chi^2 = 6.18$, $p < .05$). In terms of interest in physics, however, the conventional underachievers revealed significantly higher manifestations ($M = 1.02$, $SD = 0.83$) than the CogAct underachievers ($M = 0.06$, $SD = 0.11$; $\chi^2 = 6.44$, $p < .05$).

The female high achievers with CogAct instruction showed a significantly higher interest in physics ($M = 2.70$, $SD = 0.51$) than those with conventional instruction ($M = 1.42$, $SD = 0.55$; $\chi^2 = 28.89$, $p < .0001$) and a significantly higher self-concept in physics ($M = 3.50$, $SD = 0.25$) than the female high achievers having received conventional instruction ($M = 2.33$, $SD = 0.54$; $\chi^2 = 33.29$, $p < .0001$). Further, they used efficient learning strategies ($M = 1.91$, $SD = 0.22$) more often than their counterparts ($M = 1.24$, $SD = 0.51$; $\chi^2 = 19.08$, $p < .001$). Finally, the high achieving girls with CogAct instruction also indicated having less physics anxiety ($M = 0.23$, $SD = 0.16$) than high achieving girls with conventional instruction ($M = 1.08$, $SD = 0.48$; $\chi^2 = 13.83$, $p < .01$). All other comparisons were not significant (all χ^2 s ≤ 5.33 , all $ps \geq .07$).

Additional analysis and conclusion. Based on these findings, we further explored including the interaction term between condition and intelligence as additional predictor into the regression models that examined the general effectiveness of CogAct instruction. The interaction did not reach significance, neither for posttest conceptual understanding ($b = 0.00$, $LL\ p = .96$) and quantitative problem solving ($b = 0.02$, $LL\ p = .79$) nor for follow-up conceptual understanding ($b = -0.05$, $LL\ p = .47$) and quantitative problem solving ($b = 0.04$, $LL\ p = .65$). This outcome invited two conclusions: First, it indicated the potential of additional analyses, such as LPA, to detect intervention effects on particular groups of students. Second, there was no general disadvantage of CogAct instruction for less intelligent students – i.e., no aptitude-treatment interaction. To further illustrate this statement, Table 5.4 lists the condition-specific means on posttest and follow-up conceptual understanding and quantitative problem solving for students with below average and students with above average intelligence (i.e., median split). In brief, the descriptive statistics in Table 5.4 show that students consistently achieved higher scores in the CogAct condition than in the

conventional condition, irrespective of their intellectual potential (again, with the exception of follow-up quantitative problem solving where posttest scaffolding turned out to be crucial). Taken together, the results of the present study provided evidence for general effects of CogAct physics instruction focusing on conceptual understanding. In addition, the findings suggested special benefits for groups of female and male high potential students.

Table 5.4

Condition-Specific Means and Standard Deviations of Post and Follow-up Conceptual Understanding and Quantitative Problem Solving as a Function of Intelligence

Variables	Instructional Condition			
	CogAct		Conventional	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Below average intelligence				
Prior conceptual understanding	2.65	1.43	2.65	1.72
Post				
Conceptual understanding	5.81	2.43	5.11	2.20
Quantitative problem solving	3.45	2.41	3.09	3.01
Follow-up				
Conceptual understanding	5.21	3.22	4.64	1.73
Quantitative problem solving	2.64	2.43	3.77	2.80
Above average intelligence				
Prior conceptual understanding	3.25	1.72	3.03	1.42
Post				
Conceptual understanding	7.69	3.25	5.95	2.26
Quantitative problem solving	6.26	3.47	5.02	2.94
Follow-up				
Conceptual understanding	6.31	2.98	5.62	2.55
Quantitative problem solving	3.79	3.51	4.02	2.61

Discussion

In this study, we investigated the classroom potential of a whole unit of CogAct physics instruction that realizes a number of well-evaluated instructional elements focusing on the development of conceptual understanding. Promising physics instruction in theory does not

necessarily result in effective classroom practice. It is far from certain that instructional programs implemented in real school life really achieve the desired effects. So does CogAct instruction lead to mean learning advantages compared to conventional instruction on a conceptual as well as a procedural knowledge measure? Do CogAct students need a conceptual scaffold to access problem solving procedures? Is CogAct instruction motivating? Can it tackle girls' physics underachievement? In the following sections, we try to answer these four research questions based on the present results and derive general implications for physics instruction and assessment.

Conceptual Instruction with the CogAct Teaching Unit is Generally Effective

Immediately after the 18 lessons, students having received CogAct instruction indeed outperformed students with instruction as always in terms of their conceptual understanding in basic Newtonian mechanics. Our conceptual understanding measure required the students to transfer the content learnt to new problem contexts. The finding that CogAct students were better able to transfer what they had learnt can be seen as indicator of deep conceptual understanding and the development of flexible knowledge (e.g., Brown, 1989). Also three months after the instruction, CogAct students exceeded conventional students regarding their conceptual understanding.

In line with previous research, instruction focusing on conceptual understanding did not only promote the development of deep conceptual knowledge but was also especially beneficial in terms of related procedural knowledge (e.g., Freeman et al., 2014; Hake, 1998; Thacker et al., 1994). So CogAct students showed a significantly better quantitative problem solving performance than conventional students at the posttest. This is a crucial finding, bearing in mind that, in CogAct instruction, time was spent predominantly on developing a conceptual understanding of mechanics contents whereas considerably less time was devoted to practicing quantitative problem solving. Immediately after the CogAct instruction, students successfully managed to solve quantitative physics problems primarily based on their understanding of the underlying concepts – without intense practice (c.f. Chi et al., 1981; Hardiman et al., 1989; Heyworth, 1999; Taconis et al., 2001). CogAct students outperformed conventional students in quantitative problem solving, although, in the conventional

instruction condition, teachers tended to spend more time on practicing quantitative problem solving, corresponding to previous research on physics instruction (see Langer Tesfaye & White, 2012; Seidel et al., 2006; Taconis, 1995). However, in this study, no exact measure allowing for a quantification of time spent on problem solving in each class was implemented. Therefore, we can only speak about tendencies based on the communication with the teachers and classroom observations.

Importantly, any advantage of CogAct students in terms of quantitative problem solving had disappeared at the follow-up test three months after instruction. After the uncontrolled time of three months in between, it seems that some CogAct learners could no longer access the formalisms and procedures linked to their still existing conceptual knowledge in Newtonian mechanics. In the following section, this finding is discussed in more detail.

Fostering Connections between Concepts and Procedures Facilitates Delayed Problem Solving

We did not expect students to automatically think about underlying concepts when trying to solve quantitative physics problems but to apply problem solving routines triggered by certain cues in the problem context (c.f. Gick, 1986). Therefore, half of the students within each classroom received an additional scaffold together with each quantitative problem, prompting the students to think about the physics terms and principles that have to be considered in this problem before they start calculating. We hypothesized that this scaffold should be more beneficial for CogAct students, helping them to access the problem solving procedure via their conceptual knowledge, than for conventional learners who may have developed a less elaborated conceptual knowledge base that less strongly connects to quantitative problem solving procedures. Based on the results of the study, this assumption has to be reconsidered. First, CogAct students did not profit from the scaffold at posttest in terms of posttest quantitative problem solving. Independent of the provision of the posttest scaffold, they outperformed conventional students with regard to posttest quantitative problem solving, suggesting that they were able to successfully solve quantitative physics problems primarily based on their understanding of the underlying concepts. Hence, immediately after the CogAct instruction they did not need any help to access problem solving procedures via conceptual knowledge. Second, the follow-up scaffold was also not

effective to help CogAct students to access problem solving procedures at the follow-up test. Independent of the provision of the follow-up scaffold, CogAct and conventional students did not differ significantly in terms of their follow-up quantitative problem solving performance. Therefore, scaffolding the concept-guided activation of related procedural knowledge three months after the instruction seems to be too late to take effect. Maybe the initial connections between the respective conceptual and procedural knowledge structures (corresponding to the explicit production rules as described by Anderson & Schunn, 2000 or Chi et al., 1981) were too weak to remain accessible after a time period of three months. Finally, however, the posttest scaffold proved to be effective to particularly support the CogAct students' follow-up quantitative problem solving. Consequently, CogAct students managed to perform on a high level also at follow-up quantitative problem solving if they had received the opportunity to consolidate the connection between their conceptual knowledge and related problem solving procedures at a time when both concepts, procedures, and the connections between the two were still easily accessible (i.e., at posttest). Comparable to self-explanations, for instance, the scaffolding may have prompted the CogAct students to actively elaborate the link between their conceptual understanding and the quantitative problem at hand constituting an important learning opportunity itself. Hence, scaffolding apparently was not effective because it supported CogAct students to access procedures by activating their conceptual knowledge but because it helped to strengthen relevant cognitive structures at a critical point in time. In the conventional instruction condition, conceptual knowledge and the connection between conceptual and procedural knowledge was addressed at least less explicitly than in the CogAct instruction. There is reason to assume that students attending conventional instruction indeed learnt procedures without encoding their direct relatedness to the respective concepts. This could be one explanation why these learners did not profit at all from the posttest conceptual scaffolding in terms of their follow-up quantitative problem solving.

The present findings suggest that students who had learnt introductory Newtonian mechanics focusing on conceptual understanding and had received scaffolding at the quantitative problem solving posttest, performed better on the same test three months later than those who had not received this scaffolding. The results demonstrate the potential of combining instruction focusing on conceptual understanding with specific scaffolding methods to help students strengthen the connections between conceptual and procedural knowledge structures in the long run.

There were no effects of posttest scaffolding on conceptual understanding at follow-up testing. Independent of the provision of the scaffold, CogAct students outperformed conventional students on follow-up conceptual understanding. Although the posttest scaffold had an effect on the CogAct students' follow-up problem solving performance, strengthening the connection between concepts and related procedures evidently did not affect conceptual knowledge. This finding fits the assumption that conceptual understanding resulting in the development of conceptual knowledge guides the activation of corresponding procedural knowledge (c.f. Chi et al., 1981; Dixon & Moore, 1996). Conceptual knowledge may be accessible independent of connected procedural knowledge. To access the quantitative problem solving procedures, however, connected conceptual knowledge may have to be activated.

No General Motivational Benefits of CogAct Instruction

After the CogAct instruction, students did neither show increased interest in physics, physics self-concept, and use of efficient learning strategies nor decreased learning amotivation and physics anxiety as compared to conventional instruction. These results correspond to additional information gathered after instruction. Answering three supplementary rating scale questions, the students having received CogAct instruction did not rate the last 18 lessons as more interesting, did not think that they understood more of the content, and indicated that the instruction had not stimulated more reflection about the content than the students having received conventional instruction.

These findings are surprising given the learning advantages of CogAct instruction. It seems that the CogAct students were not aware of the beneficial learning conditions provided by the CogAct instruction. In his meta-analysis on the relationship between student evaluations of teaching and learning, Clayson (2009) found a negative association between the rigor of instruction and student evaluations of teaching. He concluded that student evaluations are associated with the students' perception of the learning process. Moreover, the relationship between student evaluations and learning seems to decrease with increasing objectivity of the learning measures. Our motivational variables can be expected to partially reflect the students' perception of the learning process. Consequently, there may be a dissociation between performance outcomes (i.e., learning) and motivational outcomes when the instruction is demanding and the performance measures are objective as it is the case in

the present study. In a student sample similar to our sample, Hänze and Berger (2007), for instance, also found a dissociation between performance measures and motivational variables comparing cooperative learning to direct instruction in physics classrooms.

Then again, maybe 18 lessons are simply not enough to lead to general positive changes in characteristics that have evolved over the course of years. Furthermore, the students received no feedback regarding their performance on the study measures and grading, in general, was unaffected by the CogAct instruction. Therefore, in terms of grades and the grading process, CogAct and conventional instruction were basically comparable. Grades, however, are highly relevant for the students and substantially influence the students' academic interests, self-concepts, boredom, and future school-engagement (see Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Pekrun, Hall, Goetz, & Perry, 2014; Poorthuis et al., 2014; von Maurice, Dörfler, & Artelt, 2014). It could hence be argued that as long as grading (i.e., what kinds of questions are asked, what kinds of competences are required) and the associated feedback do not differ between conditions, academic motivation – that is reciprocally related to the most tangible feedback students get at school, grades – is unlikely to be affected.

Finally, it is also important to bear in mind that all these findings are based on rather short self-report scales that might have not been sensitive enough to capture overall changes in physics motivation as a function of instructional condition.

No Effect of CogAct Instruction on Girls' Underachievement?

We could show that, on the average, all students profited from CogAct instruction. One group of students, however, was given special attention: intelligent girls who have been shown to systematically underachieve in physics (Hofer & Stern, 2015). We examined the existence of physics underachievement in the present student sample and compared the probability to be an underachiever in both conditions. CogAct instruction had no influence on physics underachievement defined by intelligence and physics grades. Female underachievers existed in both conditions and CogAct instruction did not reduce the probability to be an underachiever. However, similar arguments as already stated to explain the lack of effects on the motivational variables can be referred to again: CogAct instruction did not include

modifying the grading process, and physics grades – together with intelligence – were used to determine physics underachievement. Yet, it could be expected that also grades should reveal learning advantages of CogAct instruction for underachieving girls (if existing), assuming that grades reflect the students' understanding of the content (but c.f. Hofer, 2015). So did underachieving female students really not profit from CogAct instruction? To clarify this point, we additionally investigated these girls' conceptual understanding, quantitative problem solving, and physics motivation, as discussed in the following.

CogAct Instruction Boosts High Potential Students' Performance on Study Measures and Motivation

The underachieving girls. Although the female underachievers did not profit from CogAct instruction in terms of their physics grades, the present findings suggest beneficial effects on study performance and motivational measures.

Performance. The female underachievers started to gain on their higher achieving peers in terms of conceptual understanding already at posttest and performed on a level comparable to the high achieving students at the follow-up test. The underachieving girls in the conventional condition showed no such development and revealed a significantly lower conceptual understanding at follow-up than their CogAct counterparts. Still, beside the boost at follow-up, the CogAct female underachievers' conceptual understanding at posttest demonstrated only slight improvements relative to their CogAct peers and the female underachievers in the conventional condition. One possible explanation for this finding may be the fact that regular exams were written immediately before or after the study's posttest. Underachieving girls had the lowest physics grades of all of the student profiles. So maybe the proximity of the official examination on the same content had negatively influenced their attention, their motivation, and, ultimately, their test performance. At the follow-up test, however, no exam on the introductory mechanics contents threatened them and drew off their attention so that they could show what they had learnt on basic Newtonian mechanics during the CogAct instruction. This interpretation is corroborated by the finding that, independent of instructional condition, the underachieving females indicated mild to moderate physics anxiety at posttest ($M = 1.10$, $SD = 0.43$), while the mean physics anxiety in the rest of the sample appeared to be low to mild ($M = 0.63$, $SD = 0.59$). Moreover, since the

underachieving girls were spread across all of the eight classrooms, specifics of the instruction in particular classrooms between the post- and follow-up test cannot explain the outstanding increase in the performance of these girls from the post- to the follow-up test. Consequently, when the performance boost is unlikely to be due to the intermediate instruction, there is reason to assume that learning had already taken place during the 18 lessons of CogAct instruction. Whether the proximity of the official exam, a not yet developed ability to completely grasp the relevant conceptual knowledge immediately after the knowledge acquisition, or a completely different factor had negatively affected their posttest performance, remain to be clarified. What we can assume with some certainty, however, is that the high performance at follow-up test that was comparable to the performance of the high achieving students can most probably be traced back to the 18 lessons of CogAct instruction that supported the female underachievers to sustainably learn and understand Newtonian mechanics concepts.

CogAct instruction did not particularly benefit the underachieving female students' performance regarding posttest quantitative problem solving. Because they possess the intellectual potential to also manage such problems, they might rather suffer from an unfavorable motivational background and, relatedly, early knowledge deficits concerning the mathematical foundations. The results of Hofer and Stern (2015), however, suggest that the female physics underachievers' grades in mathematics are less problematic than their grades in physics. Maybe they particularly dislike the convergent nature of most conventional physics problems and would prefer more realistic scenarios with more than one possible solution that allow the creative application of (formalized) conceptual knowledge (see also Zohar, 2006; Zohar & Sela, 2003). Alternatively, or perhaps additionally, the proximity of the regular exams could again be used to explain the rather low performance of these girls.

We cannot draw any conclusions regarding the underachieving female students' follow-up quantitative problem solving performance because of the scaffolding effect that would have required an additional splitting of the profile (as explicated in the chapter "Analyses on study performance measures and motivation" that is part of the subsection addressing the fourth research question in the section "Results"). Unlike with the conceptual understanding measure, we can hence only speculate about effects of CogAct instruction on quantitative problem solving that had been indistinguishable at posttest but might have become apparent at follow-up test.

The advantage of CogAct instruction for underachieving girls was particularly reflected in the follow-up measure of conceptual understanding. This finding corresponds to the absence of effects of CogAct instruction on the general existence of underachievement and the probability to be an underachiever in physics, since physics grades were primarily based on examinations around posttest that included quantitative problem solving. Yet, independent of instructional condition, even the underachieving girls' posttest quantitative problem solving performance was not as low as their physics grades would predict. It seems that the standardized tests applied in this study enable a purer assessment of the students' translated intellectual potential than grades do. Grades, by contrast, may capture additional components such as self-discipline and values that can be considered important determinants of (school) achievement (e.g., Duckworth & Seligman, 2005; Steinmayr & Spinath, 2009), just as there may be illegitimate biases in teachers' grades (e.g., de Boer, Bosker, & van der Werf, 2010; Goddard Spear, 1984; Hofer, 2015; Read, Francis, & Robson, 2005). So what cognitive, motivational, affective, and external factors can explain the underachieving girls' low achievement in grade-relevant examinations? The present results once more stress the importance of further investigating and critically reconsidering the nature of school grades.

Motivation. While there were no general motivational benefits of CogAct instruction in the overall sample, we also looked for profile-specific differences between CogAct and conventional instruction regarding the motivational variables. Again, the proximity of the official examination could be expected to negatively influence the underachieving girls' motivation at posttest. Yet, within the female underachievers, CogAct instruction in fact positively affected two of the five motivational variables assessed. Hence, the CogAct methods included in the CogAct teaching unit seemed to have stimulated underachieving female students to apply efficient learning strategies, encompassing elaboration, organization, and control strategies, significantly more often than underachieving female students having received conventional instruction. This positive effect may be attributed to the repeated exposure to and application of methods like self-explanations, metacognitive questioning, or mental tools that learners can adopt and use as learning strategies on their own (see Vosniadou et al., 2001; White & Frederiksen, 1998; Zepeda et al., 2015). The intelligent underachieving students might have recognized the value of these learning strategies and the fit between these strategies and their own need to learn with understanding. Moreover, the underachieving girls' learning amotivation, their aversion to engage in physics lessons, was significantly lower after CogAct instruction than after conventional instruction. It hence

seems that they experienced physics instruction that was based on CogAct principles and focused on conceptual understanding as less tedious and frustrating (see Deslauriers et al., 2011; Hart, 1996; Kiemer et al., 2015; Zepeda et al., 2015; Zohar & Sela, 2003). The underachieving girls' interest in physics, however, was extremely low in the CogAct condition and even significantly lower than the low physics interest of the underachieving girls in the conventional condition.

Conclusion. To sum up, these findings suggest that CogAct instruction promoted female underachievers to adjust their learning strategies and better cope with and engage in physics instruction. Yet, it did not increase their general interest in physics as a discipline. These girls are hence unlikely to opt for a career in physics-related professions and invest their high intellectual potential in this field of work. However, it is neither realistic nor desirable to convert all intelligent females to STEM disciplines (c.f. Ceci & Williams, 2011; Wang, Eccles, & Kenny, 2013). In the case of the underachieving girls, the perhaps more important objective is to help them to make the most of their school physics experiences and to invest their intellectual potential at least in parts in learning physics. Overall, the results regarding the motivational variables together with the significant boost in conceptual understanding at follow-up are definitely promising.

The high achieving girls. Besides the potential of CogAct instruction to boost the underachieving girls' conceptual understanding and partly increase their motivation, our analyses revealed that one group of students profited even more from CogAct instruction: female high achievers.

Performance. CogAct instruction had immediate and delayed effects on their performance that was considerably enhanced in terms of both conceptual understanding and quantitative problem solving. They generally caught up with the high achieving males and even significantly exceeded the males with regard to posttest quantitative problem solving. The high achieving girls in the CogAct condition also generally outperformed the high achieving girls in the conventional condition. Although the female high achievers in the CogAct condition lost some of their conceptual understanding demonstrated at posttest after the three months of uncontrolled instruction at follow-up, they still outperformed their conventional counterparts and kept up with the male high achievers. CogAct instruction enabled these girls to acquire physics literacy that goes beyond the competences needed to perform well in grading situations.

Motivation. Regarding the motivational variables, the high achieving girls with CogAct instruction as compared to conventional instruction showed more beneficial manifestations on four of the five measures included in the study. Accordingly, they were more interested in physics, reported a higher self-concept in physics, used efficient learning strategies more often, and showed less physics anxiety than their conventional counterparts. Taken together, the findings clearly suggest marked advantages of CogAct physics instruction for intelligent girls who also show high achievement in terms of physics grades. There is broad evidence that even girls who perform well in school physics doubt their competence and ability to succeed in physics (e.g., Häussler & Hoffmann, 2000, 2002; Jansen et al., 2014; Seidel, 2006). CogAct instruction and the associated methods emphasizing real-world references, active communication, student authorship, autonomous thinking, conceptual understanding, and the joint construction of comprehensible knowledge seem to enable these girls to engage in physics more intensively. This experience in turn could be expected to encourage their competence beliefs and likewise decrease their physics anxiety. Focusing on the description of real-world phenomena and working on the underlying concepts may further have raised their interest in physics as a discipline that has more to offer than memorizing and applying formulae. CogAct instruction moreover equipped high achieving female students with learning strategies that support deep conceptual understanding in the long run.

Conclusion. Identifying and installing effective instructional programs at high school level can be considered a major objective in the endeavor to reduce the gender gap in physics. A growing body of work emphasizes the years at high school as a crucial time to consolidate gender differences in achievement, engagement, interest, and participation in STEM disciplines (see e.g., Ceci, Ginther, Kahn, & Williams, 2014; Ceci, Williams, & Barnett, 2009; Halpern, 2014). The present results are promising in view of the gender gap in physics. Instructional programs that focus on conceptual understanding and involve CogAct methods may have the potential to get some intelligent girls enthusiastic about this field of work or, at least, to eliminate some girls' aversion to physics. Our results are in line with findings indicating beneficial effects of interactive engagement methods (including, for instance, group discussions, frequent feedback, and activities fostering understanding) in university physics courses for both genders but particularly for girls (Lorenzo et al., 2006; see also Zohar & Sela, 2003). Importantly, the physics instruction implemented in the present study did not benefit intelligent females at the expense of the male students. What was advantageous to girls was also advantageous to boys, as addressed in the following.

The high achieving boys. Although CogAct instruction had no influence on the high achieving boys' generally strong motivational background (c.f. Adams et al., 2006; Bryant et al., 2013; Debacker & Nelson, 2000; Lubinski & Benbow, 1992; P. Murphy & Whitelegg, 2006b; Organisation for Economic Co-operation and Development, 2009; Osborne et al., 2003), it boosted the male high achievers' conceptual understanding immediately after the 18 lessons to a level significantly higher than that of the conventional high achieving boys. The male high achievers' performance in solving quantitative problems seems to be no function of instruction but on a high level in general. Conceptual understanding, however, that is usually not explicitly addressed in conventional physics lessons can still be boosted by adequate instruction. At follow-up, after the three months of uncontrolled instruction, the CogAct high achieving boys' performance boost in conceptual understanding had levelled out and both CogAct and conventional high achievers again performed on the usual high level. As already stated when describing the CogAct instruction's potential for intelligent females, demanding in-depth conceptual instruction seems to be a particularly good match for intelligent students, irrespective of gender.

General Implications for Physics Instruction and Assessment

The results of the present study have implications for the design of physics lessons in the future. CogAct instruction turned out to benefit all students, boost high potential students, and reduce the gap between intelligent girls and boys. Yet, high achievers of both genders and, in general, all students having received CogAct instruction on average experienced a drop in their performance from post to follow-up test. This finding stresses the importance of reactivating and taking up knowledge once learnt over and over again to prevent knowledge decline and hence suggests general and permanent modifications in the way how physics is instructed. The CogAct teaching unit in introductory Newtonian mechanics actually is developed as a first part of a whole CogAct curriculum with sequential modules that explicitly build on one another and revisit the same concepts again and again. This whole curriculum is aimed at promoting the active construction of meaningful knowledge that is augmented sequentially. We (and others, see e.g., McDermott, 1984; Rosenquist & McDermott, 1987) believe that this general focus on the active development of conceptual knowledge by means of adequate instructional methods is the key to improved learning in physics and in STEM disciplines in general (a "social-constructivist" perspective). The

success of a number of instructional programs with varying names featuring partly overlapping but also different instructional methods that, however, all share this general focus, further points to its high relevance for advancing STEM instruction (e.g., Crouch & Mazur, 2001; Freeman et al., 2014; Hake, 1998; Handelsman et al., 2004; Lorenzo et al., 2006; Thacker et al., 1994).

The female underachievers seem to have serious difficulties with traditional examinations, as reflected in their low physics grades. Since they would possess the intellectual potential to manage examination problems, they appear to experience a particular performance-hampering aversion to physics examinations that may also impinge on their capability to learn from instruction. These students may hence profit from an instructional program that transforms not only the instruction itself but also the assessment.

Aligning curriculum, instruction, and assessment should generally be considered an important objective in education (see National Research Council, 2001; Newcombe et al., 2009; Pellegrino, 2009; Shepard, 2000). Therefore, instruction has to be accompanied by (formative and summative) assessments that correspond to the way of instruction to consistently guide the students' learning. As long as the assessment and thus also the process of grading remain unmodified, the competences relevant in the grading process will always influence what is learnt from modified instruction. Consequently, to increase the positive effects of CogAct teaching on performance and motivation for all students, but particularly for the underachieving girls, the students should be informed about the general focus of both the instruction and the grade-relevant assessment. Accordingly, the measurement of conceptual understanding should be used as foundation for feedback and grading purposes, for example by means of applying the bMCT as substantial part of the examination and including quantitative problems that allow more than one possible solution and require the creative application of formalized conceptual knowledge. Knowing that a grade-relevant assessment can be mastered with conceptual understanding may motivate the underachieving girls and receiving positive feedback based on such assessments could further contribute to breaking the circle of underachievement.

Limitations

In the following section, we address some important limitations that have to be considered when interpreting the present study's results. So it has to be noted that both the conceptual understanding and the quantitative problem solving performance were rather disappointing independent of instructional condition. Immediately after instruction, the CogAct students gained less than 39% of the total conceptual understanding score and the conventional students obtained less than 33% of the total score. The results are similar for quantitative problem solving (41% and 35%) and generally even worse with regard to the follow-up test. The conceptual understanding measure might have been rather difficult for the students considering that all correct and no wrong answer alternatives had to be checked for an item to be scored one point. The quantitative problem solving test, in contrast, was not particularly difficult and explicitly designed to resemble conventional physics exercises or examinations, respectively. Yet, this test was always administered after the conceptual understanding measure. Hence, fatigue and decreased motivation could have affected the students' performance on the quantitative problem solving test. Overall, the students might not have tried as hard as possible to solve the problems given in the study context without relevant external incentives. This point links to the already discussed importance of aligning curriculum, instruction, and assessment. As soon as the CogAct teaching unit is implemented in classrooms in combination with adjusted grading processes (e.g., using conceptual understanding measures also for grading purposes), the students' performance should increase. Nevertheless, even if we had to acknowledge that students do not learn all we want them to learn in CogAct instruction, they still seem to learn significantly more than in conventional instruction.

To account for classroom dependency, statistical approaches like multi-level modeling or adjusting standard errors and χ^2 -tests of model fit to complex sampling can be applied (see Muthén & Satorra, 1995; Wu & Kwok, 2012). However, at least 30 clusters (i.e., classrooms) with at least 30 individuals per cluster are recommended to reliably perform such analyses (see Hox, 1998). In the present study, we hence accepted potential limitations resulting from disregarded dependencies in the data, but note that standard errors may be underestimated. The study design yet allowed controlling for teacher and learning environment effects since each teacher instructed two highly comparable parallel classes from the same school, offering one class CogAct instruction and the other class conventional instruction. By including a set

of control variables on the individual student level, we could further take into account additional variance in the data. While the design of the present study has several strengths that are often impossible to realize in large-scale projects, in the next step, a less controlled study that, however, involves a larger number of teachers and classrooms is planned to confirm the present findings.

Another limitation of this study can be seen in the fact that student data were gathered when the participating teachers implemented the CogAct teaching unit for the first time. Moreover, there was a rather large time gap between the teacher training and the implementation of the CogAct teaching unit for some of the teachers. In line with these unfavorable circumstances, we registered initial difficulties with attuning personal preferences to the unit's standards and a decrease in authenticity and fluency in the teaching process as compared to instruction as always. The participating teachers' conventional instruction can further be considered as above average teaching, attributable to their generally high motivation and engagement as well as the particular study context involving classroom observations and performance monitoring by means of the applied tests. In addition, it has to be born in mind that the teachers had received the CogAct teacher training before the 18 lessons of CogAct and conventional instruction. Even if not intentionally, the training may have influenced also the teachers' conventional instruction to some extent – especially in the three months after the more controlled period of the 18 lessons. To conclude, the potential of the CogAct teaching unit might have been underestimated. In the case of a repeated implementation of the CogAct teaching unit and in an even more naturalistic setting, more pronounced intervention effects may result.

In particular the analyses comparing the student profiles resulting from the LPA between CogAct and conventional instruction were partly based on small sample sizes. That restricted the kinds of analyses reasonably practicable (e.g., comparing means instead of calculating regression models including all control variables). This is why we also relied on confidence intervals and the interpretation of the overall picture. The results point to some relevant and promising differential effects of CogAct instruction on underachieving girls, on high achieving boys, and, especially, on high achieving girls. The present findings suggest investigating these effects more thoroughly.

While we examined the effects of CogAct instruction focusing on conceptual understanding in the domain of physics (or introductory Newtonian mechanics, to be precise),

the assumed rationale behind the effectiveness of such kind of instruction applies to STEM subjects dealing with complex concepts in general (see e.g., Carey, 2000; Newcombe et al., 2009; Pines & West, 1986). At the moment, however, all conclusions only hold true for physics instruction, as investigated in the present study.

Conclusion

This study provides evidence for general effects and special benefits for high potential students when conceptual understanding is fostered with CogAct instruction in physics classrooms. We want to close by deriving three broad recommendations based on the present findings:

First, physics learning in terms of both conceptual knowledge and more procedural problem solving skills can be promoted by means of instruction that focuses on the students' active development of conceptual knowledge using adequate CogAct instructional methods. Such kind of instruction also boosts the performance and benefits the motivational background of highly intelligent students enabling top-performance in physics that is independent of the students' gender.

Second, instruction focusing on conceptual understanding should be combined with specific scaffolding methods aimed at helping students to strengthen the connections between conceptual and procedural knowledge structures to increase the procedures' accessibility in the long run.

Third, to enable long-term effects on performance and motivation, 18 lessons are not enough. The focus of physics instruction has to be reconsidered more generally. By continually revisiting already acquired knowledge, physics instruction should promote the active construction of meaningful knowledge that is augmented sequentially. This should involve the alignment of instruction and assessment. What is assessed should comply with the focus of instruction. Ultimately, considering these three broad recommendations may not only help to increase the students' physics attainment and inclination but also promote outstanding performance of talented boys and girls.

References

- Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 2(1), 1–14. <http://doi.org/10.1103/PhysRevSTPER.2.010101>
- Anderson, J. R., & Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5, pp. 1–27). Mahwah, NJ: Erlbaum.
- Andersson, K. (2010). “It’s funny that we don’t see the similarities when that’s what we’re aiming for” - Visualizing and challenging teachers’ stereotypes of gender and science. *Research in Science Education*, 42(2), 281–302. <http://doi.org/10.1007/s11165-010-9200-7>
- Asparouhov, T., & Muthén, B. (2012). Auxiliary variables in mixture modeling: A 3-step approach using Mplus. *Mplus Web Notes*, 15. Retrieved from <http://statmodel2.com/examples/webnotes/webnote15.pdf>
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y.-M. (2010). Teachers’ mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180. <http://doi.org/10.3102/0002831209345157>
- Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA’s Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Bereiter, C. (1997). Situated cognition and how to overcome it. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition: Social, semiotic, and psychological perspectives* (pp. 281–300). Hillsdale, NJ: Erlbaum.
- Berthold, K., & Renkl, A. (2010). How to foster active processing of explanations in instructional communication. *Educational Psychology Review*, 22(1), 25–40. <http://doi.org/10.1007/s10648-010-9124-9>
- Boaler, J. (1997). Reclaiming school mathematics: The girls fight back. *Gender and Education*, 9(3), 285–305. <http://doi.org/10.1080/09540259721268>
- Brown, A. L. (1989). Analogical learning and transfer: What develops. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 369–412). Cambridge University Press.
- Bryant, F. B., Kastrop, H., Udo, M., Hislop, N., Shefner, R., & Mallow, J. (2013). Science anxiety, science attitudes, and constructivism: A binational study. *Journal of Science Education and Technology*, 22(4), 432–448. <http://doi.org/10.1007/s10956-012-9404-x>

- Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology*, 27(5), 777–786.
- Byun, T., & Lee, G. (2014). Why students still can't solve physics problems after solving over 2000 problems. *American Journal of Physics*, 82(9), 906–913. <http://doi.org/10.1119/1.4881606>
- Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, 21(1), 13–19. [http://doi.org/10.1016/S0193-3973\(99\)00046-5](http://doi.org/10.1016/S0193-3973(99)00046-5)
- Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3), 75–141. <http://doi.org/10.1177/1529100614541236>
- Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8), 3157–3162.
- Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, 135(2), 218–261. <http://doi.org/10.1037/a0014412>
- Cervetti, G. N., Kulikowich, J. M., & Bravo, M. A. (2015). The effects of educative curriculum materials on teachers' use of instructional strategies for English language learners in science and on student learning. *Contemporary Educational Psychology*, 40, 86–98. <http://doi.org/10.1016/j.cedpsych.2014.10.005>
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182. http://doi.org/10.1207/s15516709cog1302_1
- Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439–477.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152. http://doi.org/10.1207/s15516709cog0502_2
- Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63(1), 1–49.
- Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, 31(1), 16–30.
- Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970–977.
- Debacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *The Journal of Educational Research*, 93(4), 245–254. <http://doi.org/10.1080/00220670009598713>

- De Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1), 168–179. <http://doi.org/10.1037/a0017289>
- DeLeeuw, N., & Chi, M. T. H. (2003). Self-Explanations: Enriching a situation model or repairing a domain model? In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional Conceptual Change* (pp. 55–78). Mahwah, NJ: Lawrence Erlbaum Associates.
- Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, 332(6031), 862–864.
- Dixon, J. A., & Moore, C. F. (1996). The developmental role of intuitive principles in choosing mathematical strategies. *Developmental Psychology*, 32(2), 241–253.
- Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16(12), 939–944.
- Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, 25(6), 671–688. <http://doi.org/10.1080/09500690305016>
- Festas, I., Oliveira, A. L., Rebelo, J. A., Damião, M. H., Harris, K., & Graham, S. (2015). Professional development in self-regulated strategy development: Effects on the writing performance of eighth grade Portuguese students. *Contemporary Educational Psychology*, 40, 17–27. <http://doi.org/10.1016/j.cedpsych.2014.05.004>
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. <http://doi.org/10.1073/pnas.1319030111>
- Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., ... Pekrun, R. (2009). *PISA 2006 Skalenhandbuch. Dokumentation der Erhebungsinstrumente [PISA 2006 handbook of scales. Documentation of assessment instruments]*. Münster: Waxmann.
- Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education*, 21, 360–389.
- Gabel, D. L., Sherwood, R. D., & Enochs, L. (1984). Problem-solving skills of high school chemistry students. *Journal of Research in Science Teaching*, 21(2), 221–233. <http://doi.org/10.1002/tea.3660210212>
- Gadgil, S., Nokes-Malach, T. J., & Chi, M. T. H. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction*, 22(1), 47–61. <http://doi.org/10.1016/j.learninstruc.2011.06.002>
- Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393–408.

- Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist*, 21(1-2), 99–120. <http://doi.org/10.1080/00461520.1986.9653026>
- Goddard Spear, M. (1984). The biasing influence of pupil sex in a science marking exercise. *Research in Science & Technological Education*, 2(1), 55–60. <http://doi.org/10.1080/0263514840020107>
- Greene, J. A. (2015). Serious challenges require serious scholarship: Integrating implementation science into the scholarly discourse. *Contemporary Educational Psychology*, 40, 112–120. <http://doi.org/10.1016/j.cedpsych.2014.10.007>
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74.
- Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53(11), 1056–1065. http://doi.org/10.1007/978-3-642-20072-4_12
- Halpern, D. F. (2014). It's complicated—in fact, it's complex: Explaining the gender gap in academic achievement in science and mathematics. *Psychological Science in the Public Interest*, 15(3), 72–74. <http://doi.org/10.1177/1529100614548844>
- Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., ... others. (2004). Scientific teaching. *Science*, 304(5670), 521–522.
- Hänze, M., & Berger, R. (2007). Cooperative learning, motivational effects, and student characteristics: An experimental study comparing cooperative learning and direct instruction in 12th grade physics classes. *Learning and Instruction*, 17(1), 29–41.
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, 100(1), 105–122. <http://doi.org/10.1037/0022-0663.100.1.105>
- Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition*, 17(5), 627–638. <http://doi.org/10.3758/BF03197085>
- Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of “floating and sinking.” *Journal of Educational Psychology*, 98, 307–326. <http://doi.org/10.1037/0022-0663.98.2.307>
- Hardy, I., Schneider, M., Jonen, A., Stern, E., & Möller, K. (2005). Fostering diagrammatic reasoning in science education. *Swiss Journal of Psychology/Schweizerische Zeitschrift Für Psychologie/Revue Suisse de Psychologie*, 64(3), 207–217.

- Harris, K. R., Graham, S., & Adkins, M. (2015). Practice-based professional development and self-regulated strategy development for tier 2, at-risk writers in second grade. *Contemporary Educational Psychology*, 40, 5–16. <http://doi.org/10.1016/j.cedpsych.2014.02.003>
- Hart, C. (1996). Changing physics to suit the girls? In P. F. Murphy & C. V. Gipps (Eds.), *Equity in the classroom* (pp. 236–241). London and Washington, DC: Falmer Press and UNESCO.
- Häussler, P., & Hoffmann, L. (2000). A curricular frame for physics education: Development, comparison with students' interests, and impact on students' achievement and self-concept. *Science Education*, 84(6), 689–705. [http://doi.org/10.1002/1098-237X\(200011\)84:6<689::AID-SCE1>3.0.CO;2-L](http://doi.org/10.1002/1098-237X(200011)84:6<689::AID-SCE1>3.0.CO;2-L)
- Häussler, P., & Hoffmann, L. (2002). An intervention study to enhance girls' interest, self-concept, and achievement in physics classes. *Journal of Research in Science Teaching*, 39(9), 870–888. <http://doi.org/10.1002/tea.10048>
- Heller, K. A., Finsterwald, M., & Ziegler, A. (2010). Implicit theories of mathematics and physics teachers on gender-specific giftedness and motivation. In K. A. Heller (Ed.), *Munich studies of giftedness* (pp. 239–252). Berlin: LIT.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141–158. <http://doi.org/10.1119/1.2343497>
- Hewson, M. G., & Hewson, P. W. (1983). Effect of instruction using students' prior knowledge and conceptual change strategies on science learning. *Journal of Research in Science Teaching*, 20(8), 731–743.
- Heyworth, R. M. (1999). Procedural and conceptual knowledge of expert and novice students for the solving of a basic problem in chemistry. *International Journal of Science Education*, 21(2), 195–211. <http://doi.org/10.1080/095006999290787>
- Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.
- Hofer, S. I. (2015). *Studying gender bias in physics grading: The role of teaching experience and country*. Manuscript submitted for publication.
- Hofer, S. I., Schumacher, R., & Rubin, H. (2015). *The basic Mechanics Concept Test (bMCT): An efficient Rasch-scaled multiple choice test of fundamental conceptual understanding in Newton's mechanics*. Manuscript submitted for publication.
- Hofer, S. I., & Stern, E. (2015). *Underachievement in physics: When intelligent girls fail*. Manuscript submitted for publication.
- Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction*, 12(4), 447–465. [http://doi.org/10.1016/S0959-4752\(01\)00010-X](http://doi.org/10.1016/S0959-4752(01)00010-X)
- Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS): Construction, validity, and reliability. *Assessment*, 10(2), 178–182. <http://doi.org/10.1177/1073191103010002008>

- Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-72087-1_17
- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, 326(5958), 1410–1412.
- Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology*, 41, 13–24. <http://doi.org/10.1016/j.cedpsych.2014.11.002>
- 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. <http://doi.org/10.1016/j.lindif.2013.12.003>
- Kahle, J. B., & Lakes, M. K. (1983). The myth of equality in science classrooms. *Journal of Research in Science Teaching*, 20(2), 131–140. <http://doi.org/10.1002/tea.3660200205>
- Kamii, C., & Dominick, A. (1997). To teach or not to teach algorithms. *The Journal of Mathematical Behavior*, 16(1), 51–61.
- Kamii, C., & Dominick, A. (1998). The harmful effects of algorithms in grades 1-4. In L. J. Morrow & M. J. Kenney (Eds.), *The teaching and learning of algorithms in school mathematics: 1998 NCTM yearbook* (pp. 130–140). Reston, VA: National Council of Teachers of Mathematics.
- Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379–424. <http://doi.org/10.1080/07370000802212669>
- Kesici, S., Baloglu, M., & Deniz, M. E. (2011). Self-regulated learning strategies in relation with statistics anxiety. *Learning and Individual Differences*, 21(4), 472–477. <http://doi.org/10.1016/j.lindif.2011.02.006>
- Kiemer, K., Gröschner, A., Pehmer, A.-K., & Seidel, T. (2015). Effects of a classroom discourse intervention on teachers' practice and students' motivation to learn mathematics and science. *Learning and Instruction*, 35, 94–103. <http://doi.org/10.1016/j.learninstruc.2014.10.003>
- King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31(2), 338–368. <http://doi.org/10.3102/00028312031002338>
- Kostova, Z. (2015). Anxiety in science education. *Chemistry*, 24(1), 20–57.
- Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., ... Weiss, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente [PISA 2000: Documentation of scales]*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Langer Tesfaye, C., & White, S. (2012). *High school physics teacher preparation*. American Institute of Physics Statistical Research Center.

- Leppävirta, J., Kettunen, H., & Sihvola, A. (2011). Complex problem exercises in developing engineering students' conceptual and procedural knowledge of electromagnetics. *IEEE Transactions on Education*, 54(1), 63–66. <http://doi.org/10.1109/TE.2010.2043531>
- Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning & Education*, 2(2), 119–127. <http://doi.org/10.5465/AMLE.2003.9901663>
- Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2), 118–122. <http://doi.org/10.1119/1.2162549>
- Lubinski, D., & Benbow, C. P. (1992). Gender differences in abilities and preferences among the gifted: Implications for the math-science pipeline. *Current Directions in Psychological Science*, 1(2), 61–66. <http://doi.org/10.1111/1467-8721.ep11509746>
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. <http://doi.org/10.1111/j.1467-8624.2005.00853.x>
- Mason, L. (Ed.). (2001). Instructional practices for conceptual change in science domains [Special issue]. *Learning and Instruction*, 11(4-5).
- McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, 37(7), 24–32. <http://doi.org/10.1063/1.2916318>
- Mevarech, Z. R., & Fridkin, S. (2006). The effects of IMPROVE on mathematical knowledge, mathematical reasoning and meta-cognition. *Metacognition and Learning*, 1(1), 85–97.
- Mevarech, Z. R., & Kramarski, B. (2003). The effects of metacognitive training versus worked-out examples on students' mathematical reasoning. *British Journal of Educational Psychology*, 73(4), 449–471.
- Muller, D. A., Sharma, M. D., & Reimann, P. (2008). Raising cognitive load with linear multimedia to promote conceptual change. *Science Education*, 92(2), 278–296. <http://doi.org/10.1002/sce.20244>
- Murphy, P. K., & Cromley, J. G. (Eds.). (2015). Examining innovations: Navigating the dynamic complexities of school-based intervention research. *Contemporary Educational Psychology*, 40.
- Murphy, P., & Whitelegg, E. (2006a). Girls and physics: Continuing barriers to “belonging.” *Curriculum Journal*, 17(3), 281–305. <http://doi.org/10.1080/09585170600909753>
- Murphy, P., & Whitelegg, E. (2006b). Girls in the physics classroom: A review of the research on the participation of girls in physics. *Institute of Physics, London, UK*. Retrieved from http://oro.open.ac.uk/6499/1/Girls_and_Physics_Report.pdf
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology 1995, Vol 25*, 25, 267–316. <http://doi.org/10.2307/271070>

- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (Seventh Edition). Los Angeles, CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. (J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC: National Academy Press.
- Nenniger, P., & Nyberg, R. (1992). *Motivated Learning Strategies Questionnaire (MLSQ) (European Adaption: German, French and Swedish Versions)*. Kiel: Institut für Pädagogik der Universität Kiel.
- Newcombe, N. S., Ambady, N., Eccles, J., Gomez, L., Klahr, D., Linn, M., ... Mix, K. (2009). Psychology's role in mathematics and science education. *American Psychologist*, 64(6), 538–550. <http://doi.org/10.1037/a0014813>
- Nicholls, G. M., Wolfe, H., Besterfield-Sacre, M., Shuman, L. J., & Larpkittaworn, S. (2007). A method for identifying variables for predicting STEM enrollment. *Journal of Engineering Education*, 96(1), 33–44. <http://doi.org/10.1002/j.2168-9830.2007.tb00913.x>
- Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics - Physics Education Research*, 6(2), 1–12. <http://doi.org/10.1103/PhysRevSTPER.6.020109>
- Novick, L. R., & Hmelo, C. E. (1994). Transferring symbolic representations across nonisomorphic problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1296–1321. <http://doi.org/10.1037/0278-7393.20.6.1296>
- Ohlsson, S. (2013). Beyond evidence-based belief formation: How normative ideas have constrained conceptual change research. *Frontline Learning Research*, 1(2), 70–85. <http://doi.org/10.14786/flr.v1i2.58>
- Organisation for Economic Co-operation and Development (2006). *Assessing scientific, reading and mathematical literacy*. Paris: Organisation for Economic Co-operation and Development. Retrieved from <http://www.oecd-ilibrary.org/content/book/9789264026407-en>
- Organisation for Economic Co-operation and Development (2009). *Top of the class: High performers in science in PISA 2006*. OECD Publishing. Retrieved from <http://www.oecd-ilibrary.org/docserver/download/9809061e.pdf?expires=1394711955&id=id&accname=ocid72024074a&checksum=FBF7E1D665EB5EFDB3C197D19CE9D2D7>
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079. <http://doi.org/10.1080/0950069032000032199>

- Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology*, 106(3), 696–710. <http://doi.org/10.1037/a0036006>
- Pellegrino, J. W. (2009). The design of an assessment system for the race to the top: A learning sciences perspective on issues of growth and measurement. *Center for K–12 Assessment & Performance Management, Educational Testing Service*. Retrieved from <http://www.k12center.net/rsc/pdf/PellegrinoPresenterSession1.pdf>
- Pines, A. L., & West, L. H. (1986). Conceptual understanding and science learning: An interpretation of research within a sources-of-knowledge framework. *Science Education*, 70(5), 583–604.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*.
- Poorthuis, A. M. G., Juvonen, J., Thomaes, S., Denissen, Jaap J. A., Orobio de Castro, Bram, & van Aken, Marcel A. G. (2014). Do grades shape students' school engagement? The psychological consequences of report card grades at the beginning of secondary school. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000002>
- Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2), 211–227.
- Raven, J. C., Raven, J., & Court, J. H. (1992). *Raven's Progressive Matrices und Vocabulary Scales. Teil 4 Advanced Progressive Matrices*. (S. Bulheller & H. Häcker, Trans.). Frankfurt: Swets & Zeitlinger.
- Read, B., Francis, B., & Robson, J. (2005). Gender, “bias”, assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*, 30(3), 241–260. <http://doi.org/10.1080/02602930500063827>
- Redish, E. F., Saul, J. M., & Steinberg, R. N. (1998). Student expectations in introductory physics. *American Journal of Physics*, 66(3), 212–224. <http://doi.org/10.1119/1.18847>
- Remillard, J. T., Herbel-Eisenmann, B. A., & Lloyd, G. M. (2011). *Mathematics teachers at work: Connecting curriculum materials and classroom instruction*. Routledge.
- Resnick, L. B. (2010). Nested learning systems for the thinking curriculum. *Educational Researcher*, 39(3), 183–197.
- Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, 77(1), 1–15. <http://doi.org/10.1111/j.1467-8624.2006.00852.x>
- Rittle-Johnson, B., & Siegler, R. S. (1998). The relation between conceptual and procedural knowledge in learning mathematics: A review. In *The development of mathematical skills* (pp. 75–110). Hove, England: Psychology Press/Taylor & Francis (UK).
- Rosenquist, M. L., & McDermott, L. C. (1987). A conceptual approach to teaching kinematics. *American Journal of Physics*, 55(5), 407–415. <http://doi.org/10.1119/1.15122>

- Rost, D. H., Sparfeldt, J. R., & Schilling, S. R. (2007). *Differentielles schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten (DISK-Gitter mit SKSLF-8). Manual*. Göttingen: Hogrefe.
- Roth, W. M., Van Eijck, M., Reis, G., & Hsu, P. L. (2008). *Authentic science revisited*. Rotterdam, The Netherlands: Sense Publishers.
- Sanchez, E., Garcia-Rodicio, H., & Acuna, S. R. (2009). Are instructional explanations more effective in the context of an impasse? *Instructional Science*, 37(6), 537–563. <http://doi.org/10.1007/s11251-008-9074-5>
- Sarwar, G. S., & Trumpower, D. L. (2015). Effects of conceptual, procedural, and declarative reflection on students' structural knowledge in physics. *Educational Technology Research and Development*, 63(2), 185–201. <http://doi.org/10.1007/s11423-015-9368-7>
- Schneider, M., & Stern, E. (2010a). The cognitive perspective on learning: Ten cornerstone findings. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 69–90). Paris: OECD Publishing.
- Schneider, M., & Stern, E. (2010b). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology*, 46(1), 178–192. <http://doi.org/10.1037/a0016701>
- Schneider, M., Vamvakoussi, X., & Van Dooren, W. (2012). Conceptual change. In *Encyclopedia of the Sciences of Learning* (pp. 735–738). Retrieved from http://link.springer.com/content/pdf/10.1007/978-1-4419-1428-6_352.pdf
- Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103(4), 759–775.
- Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2), 129–184. http://doi.org/10.1207/s1532690xci2202_1
- Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology*, 99(2), 285–296.
- Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learning Environments Research*, 9(3), 253–271. <http://doi.org/10.1007/s10984-006-9012-x>
- Seidel, T., Prenzel, M., Rimmele, R., Dalehefte, I. M., Herweg, C., Kobarg, M., & Schwindt, K. (2006). Blicke auf den Physikunterricht. Ergebnisse der IPN Videostudie. *Zeitschrift für Pädagogik*, 52(6), 799–821.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 4–14.
- Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2), 209–215. <http://doi.org/10.1016/j.cognition.2012.04.005>

- Siegle, D., Rubenstein, L. D., & Mitchell, M. S. (2014). Honors students' perceptions of their high school experiences the influence of teachers on student motivation. *Gifted Child Quarterly*, 58(1), 35–50. <http://doi.org/10.1177/0016986213513496>
- Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). Cambridge University Press.
- Sinatra, G. M., & Pintrich, P. R. (2003). *Intentional Conceptual Change*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2), 115–163. http://doi.org/10.1207/s15327809jls0302_1
- Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, 40, 41–54. <http://doi.org/10.1016/j.cedpsych.2014.05.005>
- Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, 94(2), 344–355. <http://doi.org/10.1037/0022-0663.94.2.344>
- Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, 19(1), 80–90. <http://doi.org/10.1016/j.lindif.2008.05.004>
- Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction*, 13(2), 191–203.
- Taasoobshirazi, G., & Carr, M. (2008). Gender differences in science: An expertise perspective. *Educational Psychology Review*, 20(2), 149–169. <http://doi.org/10.1007/s10648-007-9067-y>
- Taconis, R. (1995). *Understanding based problem solving: Towards qualification-oriented teaching and learning in physics education*. Technische Universiteit Eindhoven.
- Taconis, R., Ferguson-Hessler, M. G. M., & Broekkamp, H. (2001). Teaching science problem solving: An overview of experimental work. *Journal of Research in Science Teaching*, 38(4), 442–468. <http://doi.org/10.1002/tea.1013>
- Thacker, B., Kim, E., Trefz, K., & Lea, S. M. (1994). Comparing problem solving performance of physics students in inquiry-based and traditional introductory physics courses. *American Journal of Physics*, 62(7), 627–633.
- UCLA: Statistical Consulting Group (2014, July 18). Mplus FAQ. How can I compute a chi-square test for nested models with the MLR or MLM estimators? Retrieved July 18, 2014, from http://www.ats.ucla.edu/stat/mplus/faq/s_b_chi2.htm

- Von Maurice, J., Dörfler, T., & Artelt, C. (2014). The relation between interests and grades: Path analyses in primary school age. *International Journal of Educational Research*, 64, 1–11. <http://doi.org/10.1016/j.ijer.2013.09.011>
- Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, 4(1), 45–69.
- Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction*, 11(4), 381–419.
- Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science*, 24(5), 770–775. <http://doi.org/10.1177/0956797612458937>
- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1), 3–118.
- Wu, J.-Y., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(1), 16–35. <http://doi.org/10.1080/10705511.2012.634703>
- Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., ... Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education*, 21(4), 335–359. <http://doi.org/10.1080/08957340802347845>
- Zepeda, C. D., Elizabeth, J., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000022>
- Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction*, 33, 131–146.
- Zohar, A. (2006). Connected knowledge in science and mathematics education. *International Journal of Science Education*, 28(13), 1579–1599. <http://doi.org/10.1080/09500690500439199>
- Zohar, A., & Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction*, 18(4), 337–353. <http://doi.org/10.1016/j.learninstruc.2007.07.001>
- Zohar, A., & Sela, D. (2003). Her physics, his physics: Gender issues in Israeli advanced placement physics classes. *International Journal of Science Education*, 25(2), 245–268. <http://doi.org/10.1080/09500690210126766>

Acknowledgements

We want to express our gratitude to Bahar Behzadi, Monica Vogel-Stalder, Conradin Beeli, André van der Graaff, and Mark Heinz. Moreover, we wish to thank Pál Molnár, Samuel Nuesch, and Sebastian Seehars for their help and Jessica Büetiger for her great assistance throughout the project. We would also like to thank Bruno Rütsche and Peter Edelsbrunner for their help with technical and software issues.

6. General Discussion

The remainder of the thesis aims at providing an overview of this work's contribution to the field of applied educational psychology. As a starting point for the discussion, the main findings are summarized and integrated by embedding them into the S(tudent)I(nstruction)A(ssessment)-Interaction-Framework. The three overarching research questions brought up in the general introduction – (1) What factors may contribute to the underachievement of some female students in secondary school physics classrooms? (2) What is the contribution of the present work regarding the gender gap in physics? (3) What is the contribution of the present work regarding the design of physics lessons? – are addressed afterwards. This closing section is focused on the overall picture drawing general conclusions concerning the immediate learning conditions in secondary school physics classrooms that go beyond the scope of each individual paper.

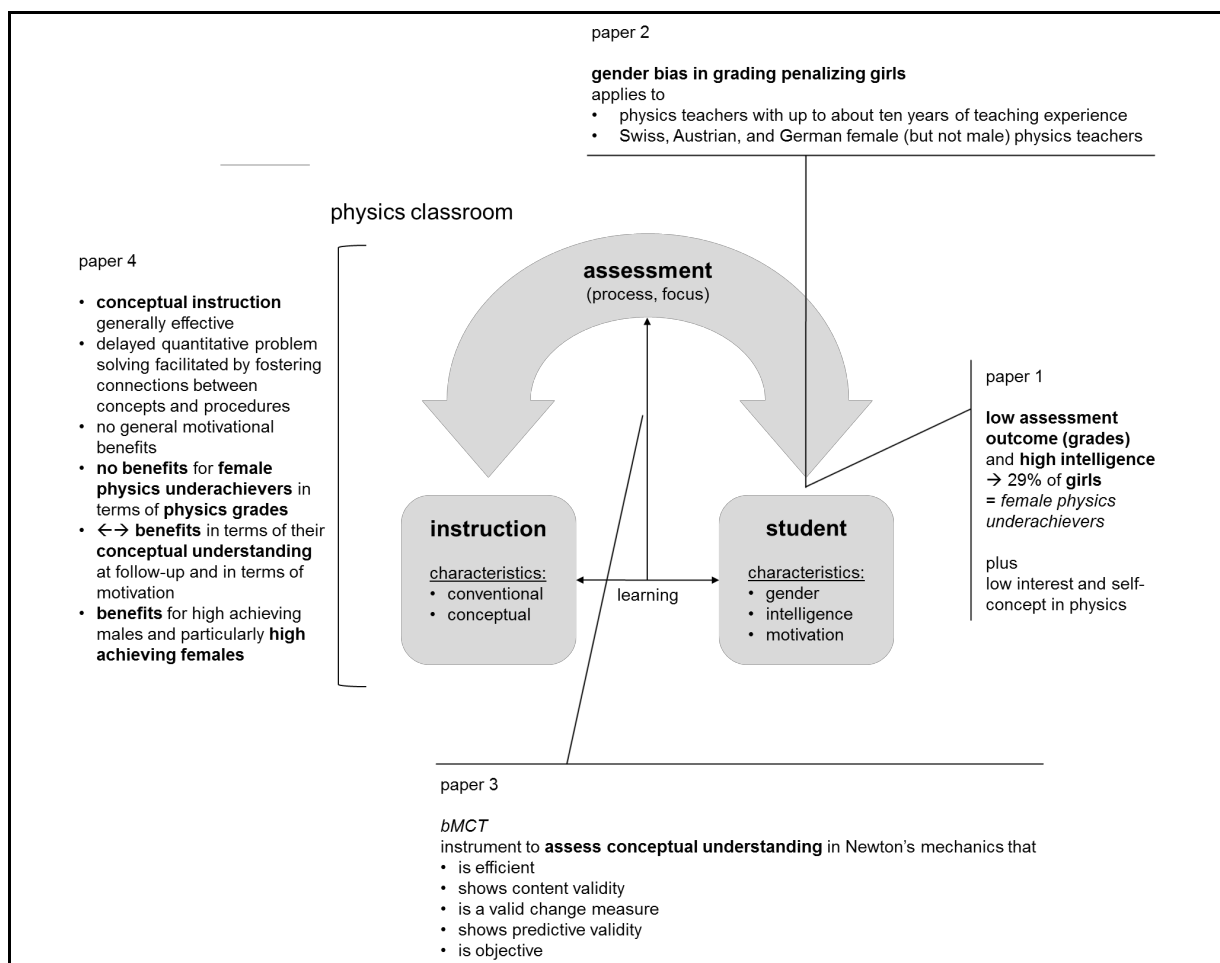


Figure 6.1. Summary of the four papers' main findings embedded into the SIA-Interaction-Framework.

Integrative Summary of the Main Findings

Figure 6.1 shows the SIA-Interaction-Framework of Physics Learning including the main results of the four papers that constitute this thesis. Focusing on the interaction between assessment and the student characteristics gender and intelligence, the first paper “Underachievement in Physics: When Intelligent Girls Fail” could corroborate the hypothesis that more intelligent female than male students underachieve in the domain of physics, i.e., receive low physics grades despite a high intellectual potential. In fact, physics underachievers were detected only among female but not among male secondary school students. The first study provided a sound rationale for considering the student characteristics gender and intelligence when investigating physics learning within the dynamics of the physics classroom.

The second paper “Studying Gender Bias in Physics Grading: The Role of Teaching Experience and Country” zoomed in on the process of assessment and its interaction with the student characteristic gender, revealing gender bias effects in grading of Swiss, Austrian, and German female physics teachers with up to about ten years of teaching experience. The cross-sectional study (paper 1) and the experimental setting in the second paper did not allow drawing conclusions about the learning process itself. Although the second paper suggested that mere assessment effects – independent of actual performance – can in fact occur and artificially decrease some girls’ physics grades, it was impossible to disentangle effects being attributable to the process and focus of the assessment, on the one hand, and deficient learning, on the other hand, when trying to partially explain some girls’ underachievement in physics.

To put it another way, is some intelligent female students’ *learning not assessed accurately* by means of grades, or rather, do some intelligent female students *learn considerably less* from a particular kind of instruction than expected? Most probably, the answer is, a bit of both. A dynamic interaction with reciprocal relations between assessment, student characteristics (female gender, high intelligence, unfavorable motivational background), and learning would also be expected based on the SIA-Interaction-Framework. Attempting to support all students, female and male, to invest their intellectual potential in understanding physics, it is important to get an idea of the processes that underlie the low achievement of some high potential students.

The third paper “The basic Mechanics Concept Test (bMCT): An Efficient Rasch-Scaled Multiple Choice Test of Fundamental Conceptual Understanding in Newton’s Mechanics” provided the basis for a more comprehensive investigation capable of addressing those open questions. The third empirical part of the thesis hence resulted in an assessment instrument (the bMCT) that can objectively, validly, and reliably measure the outcome of student learning, focusing on conceptual understanding.

This test, supplemented by some additional items, was used in the fourth paper “Fostering Conceptual Understanding with Cognitively Activating Instruction in Physics Classrooms: Evidence for General Effects and Special Benefits for High Potential Students” to monitor student learning on an alternative measure in addition to grades. Another measure, quantitative problem solving, was included that resembled physics examinations in terms of the assessment focus (quantitative questions similar to those commonly used in examinations) but differed from physics examinations with respect to the assessment process (standardized, objective, blind scoring).

So this last study, finally, investigated the effects of characteristics of the instruction on all students and, in particular, on underachieving girls. It compared conventional with cognitively activating (CogAct) conceptual instruction using alternative assessment instruments in addition to grades. This study was hence capable of looking at the whole dynamic process of learning in the physics classroom as specified by the SIA-Interaction-Framework. Conceptual instruction proved to be beneficial for all students when conceptual understanding and quantitative problem solving were assessed. It was pointless to compare the conventional and conceptual instruction classrooms in terms of physics grades because of the social referencing commonly applied on the class level when grades are assigned. Accordingly, the mean physics grade in both conditions was approximately 4.5 – as it is the case in most physics classrooms in Switzerland. Yet, if it is expected that specific groups of students profit even more from conceptual instruction than their classmates, it again makes sense to compare these students’ grades between conventional and conceptual instruction classrooms. Such special benefits were expected for underachieving female students. But these girls did not profit from conceptual instruction with respect to their physics grades. Conceptual instruction, however, could boost their performance on the conceptual understanding assessment instrument in the long run. It further seemed to have stimulated underachieving female students to apply efficient learning strategies more often than underachieving female students having received conventional instruction. Moreover, the

underachieving girls' learning amotivation was significantly lower after the CogAct conceptual instruction than after conventional instruction. Unexpectedly, this was also true for their interest in physics. While conceptual instruction proved to be advantageous for high achieving male students, too, the high achieving female students' performance and motivation experienced an additional enhancement that was exceptional.

Having summarized the main results, it is time to pick up the previously asked question about conditions underlying physics underachievement. Is some intelligent female students' *learning not assessed accurately* by means of grades or do they *learn considerably less* from a particular kind of instruction than expected or is it a combination of both? Trying to provide an answer to this question, the first of the three overarching research questions is addressed in the following section. In the course of the discussion of each of the three overarching research questions, general recommendations for future research and educational practice are derived.

What Factors May Contribute to the Underachievement of Some Female Students in Secondary School Physics Classrooms?

Referring to the literature on general scholastic underachievement, unfavorable conditions with respect to the ability self-concept, the expectancy of success, the valuation of academic goals, realistic expectations, and effective goal-related strategies, for instance, are reported to trigger a circle of underachievement. Several influencing factors are suggested to potentially affect the development of such unfavorable student characteristics. The entry in the more competitive secondary school, for example, is considered an event that may lead to a decrease in academic self-concept due to the Big Fish Little Pond Effect (see e.g., Marsh, 1987) leading to coping mechanisms such as disidentification with academics and disengagement. The enhanced academic challenge students may experience in secondary school can also increase the perceived costs (such as effort and time) of academics resulting in decreased utility, intrinsic, and attainment values concerning academics. In addition to the students' perception of school events, events at home or in the peer group can contribute to the development of unfavorable academic attitudes. So both teachers' and parents' academic

expectations, being reflected in a more or less supportive or impeding environment, shape the students motivational background (e.g., Rubenstein, Siegle, Reis, Mccoach, & Burton, 2012; Siegle, 2013; Snyder & Linnenbrink-Garcia, 2013). In the domain of physics, expectations of the social environment (home, peer group, school, or/and culture) concerning gender and STEM can be regarded as a particularly important factor favoring the emergence of those unfavorable student characteristics that may lead to some female students' underachievement (see Heller, Finsterwald, & Ziegler, 2010; Kessels, Rau, & Hannover, 2006; Leslie, Cimpian, Meyer, & Freeland, 2015; Nosek et al., 2009).

This section, however, is not meant to fathom the primary reasons explaining why some intelligent girls at first underachieve in physics. This would go beyond the scope of this thesis. Yet, within the limits of the present work, there are in fact some data available that allow drawing conclusions about conditions underlying physics underachievement as far as the interaction between the student, instruction, and assessment is concerned. Consequently, while this thesis cannot explain why physics underachievement develops in the first place, it can provide important insights how to break the circle of underachievement targeting the physics classroom. The following considerations are primarily based on the findings of the fourth paper.

Let's start with considering the interaction between the student (i.e., the underachieving girl) and characteristics of the instruction (i.e., conventional vs. conceptual instruction). The first paper indicates that female physics underachievers suffer from an exceptionally low self-concept in physics and are not interested in physics. Whereas these two motivational variables were not enhanced by instruction focusing on conceptual understanding, the underachieving girls' use of efficient learning strategies and their learning amotivation indeed were positively affected by conceptual instruction, as reported in the fourth paper. While self-concept and interest in physics are related to the students' perception of the domain of school physics on a more general level, learning amotivation and the usage of learning strategies more directly connect to the process of learning from instruction. An improvement on these two variables contingent on the instructional condition could hence be considered a direct indicator of an enhanced willingness and increased readiness potential to learn. These considerations are backed up by the results regarding the conceptual understanding assessment. As evidenced by the high conceptual understanding that female underachievers in the CogAct condition demonstrated at the follow-up test, with adequate instruction, underachieving girls definitely seem to be capable of learning and understanding Newtonian

mechanics on a level comparable with the high achieving students. In conventional instruction, by contrast, the underachieving girls are unable to catch up with the other highly intelligent learners. Or to put it another way, they do learn considerably less than expected from conventional instruction, but they are willing to learn physics as long as the instruction fits their needs. The characteristics of the instruction can make a difference.

Now, how important is the assessment in this process? Is some intelligent female students' learning not assessed accurately by means of grades? The high conceptual understanding of the underachieving girls having attended CogAct instruction became apparent at the follow-up test but not at the posttest, when the regular exams took place. At posttest, hence, not only grades but also the conceptual understanding assessment suggested that these girls had learnt less than expected during the 18 lessons of physics instruction. The quantitative problem solving assessment at posttest pointed towards this conclusion, too. In the fourth paper, the preferred explanation provided for this finding was that the official examination may have overshadowed the other two assessments that were of considerably less relevance for the students. However, this assumption remains to be tested.

Although, for reasons still unknown, there was no striking dissociation between physics grades and the study performance measures at posttest, a yet slightly different picture of physics achievement emerges dependent on the kind of assessment considered, as shown in the following. To define the profile of physics underachievers, physics grades were regarded as the most appropriate measure (see paper 1). Based on the student profiles defined in that way, Figure 6.2 visualizes the female underachievers' z-standardized average achievement at posttest as reflected by each of the three assessment outcomes (i.e., physics grades, quantitative problem solving score, and conceptual understanding score). The z-standardization was done on the whole student sample, independent of instructional condition and student profile. Hence, the z-standardized estimates allow comparing the ranking of the underachieving females relative to the other students contingent on the type of assessment.

Even though all three measures are meant to reflect the underachieving girls' knowledge acquisition during the 18 lessons of physics instruction, Figure 6.2 shows that, across both conditions, the female underachievers were ranked more than one standard deviation below the mean in terms of physics grades, but demonstrated average performance in terms of quantitative problem solving and average to slightly below average performance in terms of conceptual understanding. In both instructional conditions, the underachieving girls' physics

grades were more extreme than their quantitative problem solving performance and conceptual understanding would suggest. Besides, this illustration in fact reveals a small advantage of underachieving girls having attended CogAct instruction also with respect to posttest conceptual understanding that did differ significantly from average conceptual understanding in the conventional condition but not in the CogAct condition.

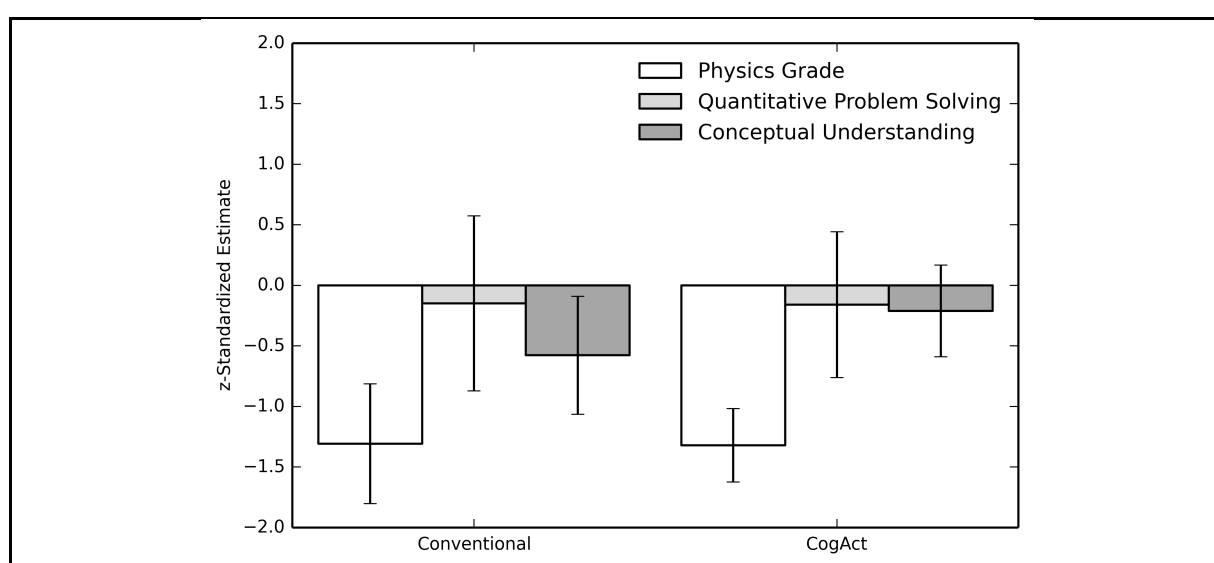


Figure 6.2. Female underachievers' z-standardized average achievement at posttest in terms of physics grades, the quantitative problem solving score, and the conceptual understanding score as a function of instructional condition. Error bars represent the 95% confidence intervals.

Although the underachieving females' high conceptual understanding of the contents taught in the course of the 18 CogAct lessons (as revealed at the follow-up test) did not show up in any of the three assessments at posttest, the physics grades appeared to be disproportionately low, independent of instructional condition. The conceptual understanding and quantitative problem solving measures that were comparable in terms of the assessment process (standardized, objective, blind scoring) but differed in terms of the assessment focus provided a similar picture regarding the underachieving students' performance. With respect to the assessments' focus, the quantitative problem solving test was designed to resemble common physics exams that can be assumed to primarily determine physics grades, supplemented by students' oral contributions. Consequently, a rather quantitative focus is unlikely to be the main reason for the extreme evaluation of the underachieving girls'

performance as reflected in the physics grades. The process of grading could hence be considered crucial in the context of physics underachievement.

Maybe the underachieving girls particularly trigger a gender bias in their teachers' grading (c.f. paper 2). Maybe the underachieving girls have severe, performance-hampering difficulties with grade-relevant exam situations in physics. Maybe it is the combination of knowing that an assessment focuses on quantitative problem solving and is of high relevance. A closer look at the focus and particularly the process of assessment in the future may shed more light on these issues.

Coming back to the previously asked question, whether female underachievers' *learning* is *not assessed accurately* by means of grades or whether they *learn considerably less* from a particular kind of instruction than expected or whether it is a combination of both, it appears that the last alternative has the greatest potential. Conventional physics instruction and the common assessment of the students' performance resulting in grades have been identified as factors that may contribute to the underachievement of some female students in secondary school physics classrooms. Deriving recommendations for educational practice, instruction that focusses on conceptual understanding and supports the active construction of knowledge in combination with modified assessments that ensure objectivity and can be mastered with conceptual understanding may tackle girls' underachievement in physics (see paper 4).

What is the Contribution of the Present Work Regarding the Gender Gap in Physics?

Speaking about the gender gap in physics in the context of the second overarching research question, the emphasis is on the gender gap with respect to achievement and motivation that can be addressed referring to the present results. Gender disparities are a recurring topic both in the papers constituting this thesis and the broad literature on science education (e.g., Deary, Strand, Smith, & Fernandes, 2007; Organisation for Economic Co-operation and Development, 2009; Taasoobshirazi & Carr, 2008). Can the findings of this thesis contribute towards reducing the gender gap in physics? Three main lines of argumentation based on the research conducted within the framework of this thesis are presented in the following.

First of all, everything that helps female physics underachievers also works against the gender gap in physics. Referring to the first paper, if all underachieving girls performed in accord with their intellectual potential, the portion of physics high achievers would be comparable between female and male students.

Second, this thesis provided evidence for the existence of a gender bias in some teachers' grading (see paper 2). This finding, taken by itself, suggests different kinds of interventions such as reducing the perceived ambiguity and cognitive overload of teachers in judgment situations, providing more structure and standardization in the assessment process, supporting (beginning) teachers in monitoring their (socio-)cognitive processes when student achievement is evaluated, or sensitizing physics teachers to the problem of gender bias in grading and gender-STEM stereotypes in general. By reducing gender bias in teachers' evaluations, the gender gap, as evidenced by teacher-dependent performance measures, should likewise diminish to some extent. Moreover, receiving, on average, more positive (or less negative) feedback on their assessed performance may positively affect the girls' physics related motivation and, in turn, future performance (e.g., Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Poorthuis et al., 2014; von Maurice, Dörfler, & Artelt, 2014).

More generally speaking, the results of this thesis stress the relevance of gender-fair assessment when trying to close the gender gap in physics. Assessments can be tested for gender fairness and can be explicitly designed to enable impartial performance measurement, not only with respect to the objectivity of the assessment process but also with respect to the functioning of different items or questions in the assessment instrument (as done in the context of the development of the bMCT, described in the third paper).

As a side note, when referring to a gender bias in grading to the detriment of girls or in favor of boys, this is meant to indicate a discrepancy. It is not meant to indicate a certain deviation from an accurate judgment because, based on the experimental approach applied in the second study, it is impossible to tell whether girls are underrated or boys overrated or both. The experimental approach allows the detection of differences in teachers' judgments as a function of student gender. This approach does not permit conclusions about the accuracy of these judgments (see also Jussim & Zanna, 2005). However, as long as such conclusions are not intended, the only major point of criticism concerning this approach may be seen in the misleading use of the term 'bias'. So it is emphasized that bias does not refer to

a systematic deviation from an accurate judgment, but is simply meant to indicate a systematic variation in teachers' judgments as a function of the experimental variation of a stereotyped characteristic. Nevertheless, the fourth study indeed provides some evidence that physics grades may in fact be biased to the detriment of girls. While Spearman's rank-order correlation between intelligence and physics grades was not significantly different from zero for female students ($r_s = -.00, p = .97$), it reached significance for male students ($r_s = .29, p < .05$). Importantly, the parallel study performance measures of conceptual understanding and quantitative problem solving did significantly correlate with the students' intelligence irrespective of gender (all $r_s \geq .24$, all $ps < .05$). So the correlation between performance and intelligence that was found for the standardized, highly objective study performance measures was also reflected in the correlation between the male students' physics grades and intelligence. Consequently, only the female students' physics grades behaved differently suggesting that a gender bias in grading may rather result from an intelligence-independent underrating of girls than an overrating of boys. Most probably, however, a bias in both directions is possible. According to the first study, some boys (the male overachievers) managed to get considerably better grades in physics than their relatively low intellectual potential would predict. Future research may specify what conditions, or which specific student characteristics, can trigger which kinds of biases and deduce what interventions have to focus on. Yet, it is important to keep in mind that teachers' bias in grading is likely to explain only a small portion of existing gender differences in physics grades (see paper 2). Nevertheless, such biases have to be taken seriously, not least because they may be considered one symptom of a generally suboptimal learning environment girls have to face in the physics classroom.

The finding that gender-STEM stereotypes seem to influence some teachers' grading suggests the existence of biases in physics teachers' behavior on a more general level. Adding to the results of the second paper, there is in fact conclusive evidence that physics teachers behave differently when interacting with female and male students. On average, boys receive more attention and are more often verbally addressed, instruction in general concentrates on the boys in the physics classroom, physics teachers are more often calling on boys than girls, approving of challenging remarks from boys but not from girls, putting more demanding questions to boys than girls, and favor the boys' approach to science over the girls' approach (e.g., Andersson, 2010; Hoffmann, 2002; McCullough, 2002; Taasobshirazi & Carr, 2008). Such unfavorable conditions might be especially harmful since girls seem to

attach more importance to pleasing the teacher in physics than boys do (Debacker & Nelson, 2000). Also the second study itself points to the possibility of more general effects of gender-STEM stereotypes on the teaching process. In this study, a student's answer to one conceptual question that could also be considered a good proxy for a student's oral classroom contribution was used as the judgment situation. The answer represented average student performance and was neither completely wrong nor absolutely correct, leaving room for interpretation. The student answer could hence be considered to reflect an intermediate state in the student's knowledge development process with some correct elements and some elements that still need to be restructured or even abandoned. A bias favoring boys (or penalizing girls, respectively) may correspond to different ways of interpreting the student's answer. Hence, such an answer originating from a boy may be interpreted as indicating a promising step on the way to full understanding (in the sense of a benefit of the doubt) resulting in supportive instructional actions. On the contrary, such an answer originating from a girl may be interpreted as indicating profound misconceptions being hard to overcome resulting in teachers resigning and putting less effort into supportive instructional actions. To conclude, instruction that is explicitly designed to support all students' active knowledge construction could be expected to particularly help girls by providing support that might otherwise be less available. The above paragraph links to the third point that is discussed in the remainder of this section.

Third and finally, as described in the fourth paper, CogAct instruction that focused on the active construction of conceptual knowledge did not only support female underachievers but also, and above all, female high achievers. Although female high achievers' physics grades were exceptional and hence in accordance with their high intellectual potential independent of the instructional condition, the conceptual understanding and quantitative problem solving measures still revealed a superiority of the high achieving boys in the conventional instruction condition. In the CogAct condition, however, the high achieving female students generally caught up with the high achieving males in terms of both study performance measures and even significantly exceeded the males with regard to posttest quantitative problem solving. The high achieving girls in the CogAct condition also generally outperformed the high achieving girls in the conventional condition. CogAct instruction enabled these girls to acquire physics literacy that goes beyond the competences needed to perform well in grading situations.

In addition to foster conceptual and procedural knowledge, improving especially female students' attitudes towards physics, their inclination to engage in physics, and their perceived competence to succeed in physics have to be considered important objectives of physics education in light of the gender gap (e.g., Nicholls, Wolfe, Besterfield-Sacre, Shuman, & Larпкиattaworn, 2007; Osborne, Simon, & Collins, 2003). A growing body of work emphasizes that girls' experiences at high school considerably contribute to their turning away from STEM disciplines (see e.g., Ceci, Ginther, Kahn, & Williams, 2014; Ceci, Williams, & Barnett, 2009; Halpern, 2014). There is broad evidence that, for girls, performing well in school physics does not automatically imply a high physics self-concept and positive physics-related attitudes and motivation (e.g., Häussler & Hoffmann, 2000, 2002; Jansen, Schroeders, & Lüdtke, 2014; Seidel, 2006). After having attended CogAct instruction, female high achievers did not only report a higher self-concept in physics, but were also more interested in physics, used efficient learning strategies more often, and showed less physics anxiety than their conventional counterparts. Taken together, the findings clearly suggest that the female high achievers got stimulated and inspired by CogAct instruction and the associated methods (see paper 4). The development of more intrinsic as compared to extrinsic motivation and the acquisition of flexible conceptual knowledge that is meaningful outside school and the next exam situation may pave the way for an increased number of female top performers in physics.

What is the Contribution of the Present Work Regarding the Design of Physics Lessons?

“Wenn man sich nach den Mädchen richtet, so ist es auch für die Jungen richtig; umgekehrt aber nicht.“

(Wagenschein, 1965, p. 350)

This quote from the renowned German physicist and science educator Martin Wagenschein – which can be translated as “If you orient towards the girls, it is advantageous to the boys, too; but not vice versa.” – provides a good summary of the present work's main results. Reviewing the above sections, it becomes clear that what is advantageous to underachieving girls is advantageous to girls in general. And the findings described in the

fourth paper indicate that CogAct physics instruction focusing on conceptual understanding did not benefit intelligent females at the expense of the male students, but rather boosted the male high achievers' conceptual understanding immediately after the 18 lessons to a level significantly higher than that of the conventional high achieving boys, too. Consequently, demanding in-depth conceptual instruction seems to promote outstanding performance of talented students, irrespective of their gender. On the contrary, conventional instruction and assessment (i.e., conventional physics examinations) indeed tend to benefit male students at the expense of female students. So, in conventional instruction, there are no male underachievers but male overachievers in terms of physics grades (see paper 1), there is evidence for a gender bias in grading (see paper 2), and male high achievers dominate over female high potential students (see paper 4).

As specified in the general introduction to this thesis, this work was intended to figure out what can be done to help secondary school students, both female and male, to invest their intellectual potential in understanding physics by adding new perspectives and findings that emphasize the interplay between gender, underachievement, and conceptual instruction. This thesis accordingly determined students who require particular attention, i.e., the underachieving females, and investigated what factors might contribute to and what factors might remedy some intelligent girls' underachievement in physics. The underachieving girls may be considered the most sensitive indicator revealing what should be modified in the design of physics lessons. Other student groups appear to be better able to cope with conventional physics lessons. Trying to design physics lessons that suit these girls seems to support secondary school students in general to invest their intellectual potential in understanding physics.

To put it in a nutshell, the present findings suggest that physics lessons should include instruction that focuses on the students' active development of conceptual knowledge using adequate CogAct instructional methods. Moreover, instruction and assessment should be aligned. What is assessed should comply with the focus of instruction. Importantly, the process of assessment should be reconsidered, too. So the perceived ambiguity and cognitive overload teachers may experience in judgment situations might be reduced by providing more structure and standardization in the assessment process. Teachers should be supported in monitoring their (socio-)cognitive processes when student achievement is evaluated and sensitized to the problem of gender bias in grading and gender-STEM stereotypes in general.

Overall, more importance should be attached to the design of good, gender-fair assessment instruments.

The SIA-Interaction-Framework of Physics Learning that was used in this thesis to embed the four papers and define a general frame for the research conducted may also be helpful to guide future research concerned with understanding and designing physics lessons. The three factors *student*, *instruction*, and *assessment* as well as their *interactions* can be considered to play an important role in understanding the process and the outcomes of learning in the physics classroom. Definitely, there are additional factors that can influence learning. The classroom as a whole, including characteristics such as classroom climate, classroom-level performance and intelligence or gender composition, for instance, may also interact with the three factors incorporated in the framework. Although the classroom can certainly have effects (e.g., Trautwein & Baeriswyl, 2007; Trautwein, Lüdtke, Marsh, & Nagy, 2009), in the present work, it was regarded as a factor of secondary importance that may be investigated in a next step. An investigation of this factor requires a different research design involving considerably more classrooms. Yet, the SIA-Interaction-Framework is open to amendments to guide more comprehensive research projects. The characteristics that were examined in terms of the student and the instruction may also be extended in future studies. So a broader range of student variables, including affective and personality variables or the students' gender stereotype endorsement, may be added to the analysis. While, in the present work, a rather coarse-grained level of analysis was chosen to investigate characteristics of the instruction (conceptual vs. conventional), more fine-grained analyses may follow that could, for example, contrast instructional forms like teacher-centered, self-regulated, or problem-based learning all focusing on conceptual understanding. Such kind of research can inform the design of physics lessons and complement the recommendations that are based on the findings of this thesis.

The SIA-Interaction-Framework may also be applied in teacher education and consulted by practicing teachers. The framework could be a helpful tool to stimulate thought and discussion about the three factors and, in particular, about their interactions. Being sensitized to the complex dynamics in the classroom, retrospective, instantaneous, or forward-looking meta-cognition about a physics lesson may be facilitated and individual students' learning processes may be better understood. Regarding both research and educational practice, the SIA-Interaction-Framework may be adapted to domains other than physics, too. Contingent on the specific domain, other student characteristics may be of particular interest and

different interactions may be assumed. As far as characteristics of the instruction are concerned, the rationale behind the effectiveness of conceptual instruction generally applies to, at least, all STEM subjects dealing with complex concepts (see e.g., Carey, 2000; Newcombe et al., 2009; Pines & West, 1986).

This thesis closes citing two of the most famous physicists of the last century. Albert Einstein suffered from the teaching methods and the concentration on rote learning at the expense of individual and creative thinking at the Gymnasium (see Fölsing, 1998). As a critic of the common focus of teaching at school, the following undated quote exemplarily expresses his opinion.

“The only thing that interferes with my learning is my education.”

(Albert Einstein)

Education at school, in physics and in general, should be inspiring and not detrimental to learning as it may sometimes be the case, especially for underachieving girls, in conventional physics classrooms. By studying the interplay between gender, underachievement, and conceptual instruction in secondary school physics, this thesis can contribute to the field by adding some recommendations how physics education could look like to help and not hinder secondary school students, both female and male, to invest their intellectual potential in learning.

Instruction focusing on conceptual understanding that is aligned with gender-fair assessment can be expected to be beneficial for intelligent students of both genders, corroborating the words of Martin Wagenschein stated at the beginning of this final chapter and matching a statement of another highly respected physicist of the last century, well known for his inspiring teaching:

“I think, however, that there isn’t any solution to this problem of education other than to realize that the best teaching can be done only when there is a direct individual relationship between a student and a good teacher – a situation in which the student discusses the ideas, thinks about the things, and talks about the things. It’s impossible to learn very much by simply sitting in a lecture, or even by simply doing problems that are assigned.”

(Richard P. Feynman in the Preface to *The Feynman Lectures on Physics*, June, 1963)

References

- Andersson, K. (2010). "It's funny that we don't see the similarities when that's what we're aiming for" - Visualizing and challenging teachers' stereotypes of gender and science. *Research in Science Education*, 42(2), 281–302. <http://doi.org/10.1007/s11165-010-9200-7>
- Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, 21(1), 13–19. [http://doi.org/10.1016/S0193-3973\(99\)00046-5](http://doi.org/10.1016/S0193-3973(99)00046-5)
- Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3), 75–141. <http://doi.org/10.1177/1529100614541236>
- Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, 135(2), 218–261. <http://doi.org/10.1037/a0014412>
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21. <http://doi.org/10.1016/j.intell.2006.02.001>
- Debacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *The Journal of Educational Research*, 93(4), 245–254. <http://doi.org/10.1080/00220670009598713>
- Fölsing, A. (1998). *Albert Einstein: A Biography*. (E. Osers, Trans.). New York, N.Y.: Penguin Books.
- Halpern, D. F. (2014). It's complicated—in fact, it's complex: Explaining the gender gap in academic achievement in science and mathematics. *Psychological Science in the Public Interest*, 15(3), 72–74. <http://doi.org/10.1177/1529100614548844>
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, 100(1), 105–122. <http://doi.org/10.1037/0022-0663.100.1.105>
- Häussler, P., & Hoffmann, L. (2000). A curricular frame for physics education: Development, comparison with students' interests, and impact on students' achievement and self-concept. *Science Education*, 84(6), 689–705. [http://doi.org/10.1002/1098-237X\(200011\)84:6<689::AID-SCE1>3.0.CO;2-L](http://doi.org/10.1002/1098-237X(200011)84:6<689::AID-SCE1>3.0.CO;2-L)
- Häussler, P., & Hoffmann, L. (2002). An intervention study to enhance girls' interest, self-concept, and achievement in physics classes. *Journal of Research in Science Teaching*, 39(9), 870–888. <http://doi.org/10.1002/tea.10048>

- Heller, K. A., Finsterwald, M., & Ziegler, A. (2010). Implicit theories of mathematics and physics teachers on gender-specific giftedness and motivation. In K. A. Heller (Ed.), *Munich studies of giftedness* (pp. 239–252). Berlin: LIT.
- Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction*, 12(4), 447–465. [http://doi.org/10.1016/S0959-4752\(01\)00010-X](http://doi.org/10.1016/S0959-4752(01)00010-X)
- 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. <http://doi.org/10.1016/j.lindif.2013.12.003>
- Jussim, L., & Zanna, M. P. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology*, 37, 1–93.
- Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, 76(4), 761–780. <http://doi.org/10.1348/000709905X59961>
- Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219), 262–265. <http://doi.org/10.1126/science.1261375>
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280–295. <http://doi.org/10.1037/0022-0663.79.3.280>
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. <http://doi.org/10.1111/j.1467-8624.2005.00853.x>
- McCullough, L. (2002). Women in physics: A review. *The Physics Teacher*, 40(2), 86–91. <http://doi.org/10.1119/1.1457312>
- Newcombe, N. S., Ambady, N., Eccles, J., Gomez, L., Klahr, D., Linn, M., ... Mix, K. (2009). Psychology's role in mathematics and science education. *American Psychologist*, 64(6), 538–550. <http://doi.org/10.1037/a0014813>
- Nicholls, G. M., Wolfe, H., Besterfield-Sacre, M., Shuman, L. J., & Larpkittaworn, S. (2007). A method for identifying variables for predicting STEM enrollment. *Journal of Engineering Education*, 96(1), 33–44. <http://doi.org/10.1002/j.2168-9830.2007.tb00913.x>
- Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26), 10593–10597. <http://doi.org/10.1073/pnas.0809921106>
- Organisation for Economic Co-operation and Development. (2009). *Top of the class: High performers in science in PISA 2006*. OECD Publishing. Retrieved from <http://www.oecd-ilibrary.org/docserver/download/9809061e.pdf?expires=1394711955&id=id&accname=ocid72024074a&checksum=FBF7E1D665EB5EFDB3C197D19CE9D2D7>

- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079. <http://doi.org/10.1080/0950069032000032199>
- Pines, A. L., & West, L. H. (1986). Conceptual understanding and science learning: An interpretation of research within a sources-of-knowledge framework. *Science Education*, 70(5), 583–604.
- Poorthuis, A. M. G., Juvonen, J., Thomaes, S., Denissen, Jaap J. A., Orobio de Castro, Bram, & van Aken, Marcel A. G. (2014). Do grades shape students' school engagement? The psychological consequences of report card grades at the beginning of secondary school. *Journal of Educational Psychology*, No Pagination Specified. <http://doi.org/10.1037/edu0000002>
- Rubenstein, L. D., Siegle, D., Reis, S. M., McCoach, D. B., & Burton, M. G. (2012). A complex quest: The development and research of underachievement interventions for gifted students. *Psychology in the Schools*, 49(7), 678–694. <http://doi.org/10.1002/pits.21620>
- Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learning Environments Research*, 9(3), 253–271. <http://doi.org/10.1007/s10984-006-9012-x>
- Siegle, D. (2013). *The underachieving gifted child: Recognizing, understanding, and reversing underachievement*. Waco, TX: Prufrock Press.
- Snyder, K. E., & Linnenbrink-Garcia, L. (2013). A developmental, person-centered approach to exploring multiple motivational pathways in gifted underachievement. *Educational Psychologist*, 48(4), 209–228. <http://doi.org/10.1080/00461520.2013.835597>
- Taasoobshirazi, G., & Carr, M. (2008). Gender differences in science: An expertise perspective. *Educational Psychology Review*, 20(2), 149–169. <http://doi.org/10.1007/s10648-007-9067-y>
- Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. *Zeitschrift Für Pädagogische Psychologie*, 21(2), 119–133. <http://doi.org/10.1024/1010-0652.21.2.119>
- Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology*, 101(4), 853–866.
- Von Maurice, J., Dörfler, T., & Artelt, C. (2014). The relation between interests and grades: Path analyses in primary school age. *International Journal of Educational Research*, 64, 1–11. <http://doi.org/10.1016/j.ijer.2013.09.011>
- Wagenschein, M. (1965). *Ursprüngliches Verstehen und exaktes Denken: Pädagogische Schriften*. E. Klett.

Appendix A

Motivational Scales

Interest in Physics

<i>Zurzeit macht es mir Spass, mich mit Physik-Themen zu befassen.</i>	stimme gar nicht zu	stimme eher nicht zu	stimme eher zu	stimme ganz zu
<i>Zurzeit lese ich gerne etwas über Physik.</i>	stimme gar nicht zu	stimme eher nicht zu	stimme eher zu	stimme ganz zu
<i>Zurzeit beschäftige ich mich gerne mit Physik-Problemen.</i>	stimme gar nicht zu	stimme eher nicht zu	stimme eher zu	stimme ganz zu
<i>Zurzeit bin ich interessiert, Neues in der Physik zu lernen.</i>	stimme gar nicht zu	stimme eher nicht zu	stimme eher zu	stimme ganz zu

Self-Concept in Physics

<i>Zurzeit habe ich das Gefühl, ich weiss in Physik die Antwort auf eine Frage schneller als die Anderen.</i>	trifft gar nicht zu	trifft kaum zu	trifft eher nicht zu	trifft eher zu	trifft überwiegend zu	trifft genau zu
<i>Zurzeit habe ich das Gefühl, es fällt mir in Physik leicht, Probleme zu lösen.</i>	trifft gar nicht zu	trifft kaum zu	trifft eher nicht zu	trifft eher zu	trifft überwiegend zu	trifft genau zu
<i>Zurzeit habe ich das Gefühl, ich gehöre in Physik zu den Guten.</i>	trifft gar nicht zu	trifft kaum zu	trifft eher nicht zu	trifft eher zu	trifft überwiegend zu	trifft genau zu
<i>Ich habe zurzeit ein gutes Gefühl, was meine Arbeit in Physik angeht.</i>	trifft gar nicht zu	trifft kaum zu	trifft eher nicht zu	trifft eher zu	trifft überwiegend zu	trifft genau zu

Learning Strategies in Physics

<i>Beim Lernen für den Physikunterricht versuche ich zurzeit immer wieder herauszufinden, welchen Lernstoff ich noch nicht verstehe.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik stelle ich mir die Inhalte an Beispielen vor.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik ist mir klar, was bei dem gerade behandelten Thema besonders wichtig und was eher unwichtig ist.</i>	fast nie	manchmal	oft	fast immer
<i>Wenn ich zurzeit für Physik lerne und etwas nicht verstehe, suche ich nach zusätzlicher Information, um das Problem zu klären.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik gehen mir viele Ideen zu den behandelten Themen durch den Kopf.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik wende ich den Stoff auf andere Aufgaben/Beispiele/Experimente an.</i>	fast nie	manchmal	oft	fast immer
<i>Zurzeit stelle ich mir selbst Fragen, um sicherzustellen, dass ich den Stoff, der im Physikunterricht behandelt wurde, verstanden habe.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik denke ich darüber nach, wie die Dinge im Einzelnen zusammenhängen.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik versuche ich Zusammenhänge zu sehen.</i>	fast nie	manchmal	oft	fast immer

Learning Amotivation in Physics

<i>In meinem jetzigen Unterricht in Physik habe ich keine Lust, mich zu beteiligen.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik bin ich mit meinen Gedanken woanders.</i>	fast nie	manchmal	oft	fast immer
<i>In meinem jetzigen Unterricht in Physik habe ich keine Lust, mich mit den Inhalten auseinander zu setzen.</i>	fast nie	manchmal	oft	fast immer

Physics Anxiety

Bitte beurteilen Sie die Aussage dahingehend, wie ängstlich Sie sich während der dargestellten Situation fühlen würden.


<i>Tabellen in einem Physikbuch verwenden müssen.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Einen Tag davor über eine anstehende Physikprüfung nachdenken.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Einem Lehrer zuschauen, wie er ein physikalisches Problem an der Tafel erarbeitet.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Eine Prüfung im Physikunterricht schreiben.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Viele schwierige Physikprobleme als Hausaufgabe bis zur nächsten Stunde aufbekommen.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Einem Lehrervortrag im Physikunterricht zuhören.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Einem anderen Schüler beim Erklären eines Physikproblems zuhören.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst

<i>Einen unangekündigten Test im Physikunterricht schreiben.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst
<i>Ein neues Kapitel in einem Physikbuch beginnen.</i>	niedrige Angst	ein bisschen Angst	mittelstarke Angst	ziemliche Angst	hohe Angst


Appendix B

bMCT (plus)

bMCT



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



DAS *MINT*-LERNZENTRUM DER ETH ZÜRICH

Grundfragen zur Mechanik

Vorname:

Name:

Schule:

Klasse:

Datum:

Alter: ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20 ☐ >20

Geschlecht: ☐ weiblich ☐ männlich

Wichtiger Hinweis:

Bei den folgenden Fragen können mehrere Antworten richtig sein. Kreuzen Sie alle richtigen Antworten an. Markieren Sie bitte ein deutliches Kreuz ins Kästchen: ☒

Machen Sie bitte die Kreuze mit einem Bleistift und drücken Sie fest auf, so dass man das Kreuz gut lesen kann. Wenn Sie korrigieren möchten, radieren Sie das Kreuz sauber aus.

Versuchen Sie, alle Aufgaben zu lösen. Halten Sie sich nicht zu lange bei einer einzelnen Aufgabe auf.

1. Ein volles Wasserglas steht stabil auf der Rückbank eines konstant geradeaus fahrenden Autos. Plötzlich tritt der Fahrer das Gaspedal durch und beschleunigt das Auto. Welche der folgenden Aussagen treffen zu?
- ☐ Weil sich das Glas bezüglich der Rückbank im Auto nicht bewegt, bleibt die Wasseroberfläche unverändert.
 - ☐ Das Wasser wird mit dem Auto beschleunigt, so dass etwas Wasser in Fahrtrichtung über den Rand des Glases schwappt.
 - ☐ Aufgrund der Trägheit des Wassers verändert sich die Wasseroberfläche nicht.
 - ☐ Das Wasser behält zunächst seinen vorherigen Bewegungszustand bei, so dass etwas Wasser entgegen der Fahrtrichtung über den Rand des Glases schwappt.

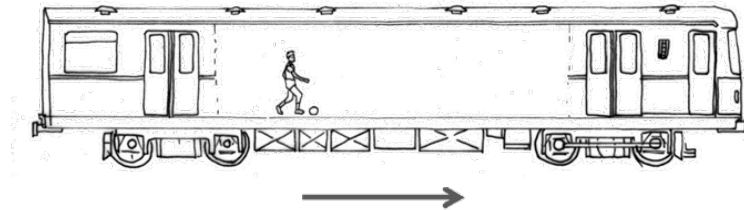
2. Ein Buch liegt vor dir auf dem Tisch. Welche der folgenden Aussagen treffen zu?
- ☐ Wie auf jeden anderen ruhenden Körper wirkt auf das Buch nur die Anziehungskraft der Erde.
 - ☐ Der Tisch stützt das Buch ab und wirkt deshalb mit einer nach oben gerichteten Kraft auf das Buch.
 - ☐ Da das Buch in Ruhe ist, kann hier überhaupt nicht mit dem Kraftbegriff argumentiert werden.
 - ☐ Auf das Buch wirkt nur die Stützkraft des Tisches, sonst würde es herunterfallen.

This item is omitted in the 11-items version of the bMCT.

3. Ein Bus fährt mit konstanter Geschwindigkeit auf horizontaler Strasse geradeaus. Welche der folgenden Aussagen treffen zu?
- ☐ Damit der Bus nicht langsamer wird, muss die Antriebskraft des Motors genau so gross sein wie der Luftwiderstand und die übrigen Reibungskräfte zusammen.
 - ☐ Damit die Geschwindigkeit konstant bleibt, muss die Antriebskraft grösser sein als der Luftwiderstand und die übrigen Reibungskräfte zusammen.
 - ☐ Damit die Geschwindigkeit nicht zunimmt, muss die Antriebskraft etwas geringer sein als der Luftwiderstand und die übrigen Reibungskräfte zusammen.
 - ☐ Die Antriebskraft ist nur zum Beschleunigen erforderlich, bei konstanter Geschwindigkeit hingegen nicht.
4. Ein Wanderer hebt einen Stein auf und geht mit 1 m/s weiter. Nach kurzer Zeit lässt er den Stein im Gehen aus 1 Meter Höhe wieder fallen. Nach $\frac{1}{2}$ Sekunde trifft der Stein wieder am Boden auf. Wo landet der Stein?
- ☐ Der Stein landet in etwa $\frac{1}{2}$ Meter hinter dem Wanderer, weil der Wanderer in $\frac{1}{2}$ Sekunde etwa $\frac{1}{2}$ Meter zurücklegt.
 - ☐ Der Stein landet in etwa neben den Füßen des Wanderers, da der Stein aufgrund seiner Trägheit seine horizontale Bewegung beibehält.
 - ☐ Weil der Stein in einem nach hinten gerichteten Bogen zu Boden fällt, landet er in etwa 1 Meter hinter dem Wanderer.
 - ☐ Da der Stein aufgrund seiner Trägheit seine horizontale Bewegung beibehält, landet er in etwa $\frac{1}{2}$ Meter vor dem Wanderer.

The positions of the items 4 and 5 are interchanged in the final version based on the final item difficulties (items sorted by difficulty in ascending order).

5. a) Ein Junge spielt im Gang des Wagens eines mit konstanter Geschwindigkeit geradeaus fahrenden Zuges mit seinem Ball. Welche Aussagen treffen zu?



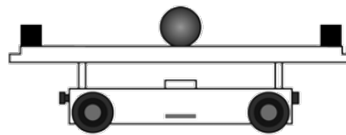
- ☐ Je nachdem, ob er den Ball aus der Mitte des Wagens in oder entgegen der Fahrtrichtung tritt, ist der Ball unterschiedlich schnell am Ende des Wagens.
- ☐ Wirft er den Ball in die Luft, muss er entweder entgegen oder in Fahrtrichtung laufen, um den Ball wieder aufzufangen.
- ☐ Der Ball verhält sich immer so, wie wenn der Zug stehen würde.
- ☐ Je nachdem, ob er den Ball in oder entgegen der Fahrtrichtung tritt, benötigt er zum Treten unterschiedlich viel Kraft.

b) Welche der folgenden Erklärungen liegt/liegen deiner/deinen Antwort(en) zugrunde?

- ☐ Sobald der Ball hochgeworfen wird, bleibt seine Bewegung hinter der des Wagens zurück. Der Grund dafür liegt in der Trägheit des Balls, die sich gegen die Bewegungsänderung stemmt.
- ☐ Wenn der Ball hochgeworfen wird, addiert sich die vertikale Bewegung vektoriell zu der horizontalen Bewegung in Fahrtrichtung. Deshalb bewegt sich der Ball vom Jungen in Fahrtrichtung weg.
- ☐ Um den Ball entgegen der Fahrtrichtung abzutreten, ist mehr Kraft erforderlich, als in Fahrtrichtung, weil man entgegen der Fahrtrichtung zusätzlich gegen die Bewegungsrichtung des Balls antreten muss.
- ☐ Aufgrund seiner Trägheit bewegt sich der Ball immer mit derselben horizontalen Geschwindigkeit wie der Wagon, sofern er nicht in oder entgegen der Fahrtrichtung getreten wird.
- ☐ Wird der Ball aus der Mitte nach vorne abgetreten, bewegt sich der vordere Teil des Wagens vom Ball weg. Wird der Ball nach hinten abgetreten, bewegt sich der hintere Teil des Wagens auf den Ball zu. Die Zeiten, bis der Ball jeweils das Ende des Wagens erreicht, sind daher verschieden.

6. Auf einem Modellwagen befindet sich eine Metallkugel, die auf einer Schiene ungehindert nach links und rechts rollen kann. An den Enden der Schiene sind Begrenzungen angebracht, die verhindern, dass die Kugel von der Schiene rollen kann.

In der Ausgangssituation ist die Kugel in der Mitte des Wagens.



Was geschieht, wenn der Wagen aus dem Stillstand nach rechts angestossen wird?

- ☐ Die Kugel rollt auf der Schiene nach links entgegen der Fahrtrichtung.
 - ☐ Die Kugel ändert ihre Position auf dem Wagen nicht.
 - ☐ Die Kugel rollt auf der Schiene nach rechts in Fahrtrichtung.
 - ☐ Die Kugel bleibt bezüglich der Tischoberfläche in etwa am selben Ort bis sie an der Begrenzung des Wagens anstösst.
7. Nachdem ein Körper angestossen wurde, gleitet er auf einer glatten Oberfläche reibungs- und luftwiderstandsfrei dahin. Welche Aussagen treffen zu?
- ☐ Der Schwung durch das Anstossen verbraucht sich mit der Zeit. Deshalb wird der Körper immer langsamer, bis er schliesslich zum Stillstand kommt.
 - ☐ Da er sich bewegt, muss auf den Körper eine Kraft in Bewegungsrichtung wirken.
 - ☐ Die Masse des Körpers wirkt der Bewegung entgegen. Je schwerer der Körper ist, desto schneller wird er zur Ruhe kommen.
 - ☐ Der Körper gleitet mit konstanter Geschwindigkeit über die Oberfläche.
 - ☐ Der Körper ändert seine Bewegung nicht, weil keine horizontale Kraft auf ihn wirkt.

8. Eine Person steht in einem ruhenden Boot und wirft mit Schwung einen grossen Stein ins Wasser hinter dem Boot. Welche Aussagen treffen zu?



- ☐ Das Boot bewegt sich in Wurfrichtung des Steins.
 - ☐ Der Stein verdrängt Wasser und dadurch schaukelt das Boot nur etwas hin und her.
 - ☐ Lässt man einen aufgeblasenen Luftballon durch die Luft zischen, so passiert im Prinzip dasselbe.
 - ☐ Das Boot bewegt sich entgegen der Wurfrichtung des Steins.
9. Die folgenden vier Abbildungen zeigen jeweils eine Kugel, die sich auf verschiedenen geneigten bzw. gekrümmten Bahnen reibungsfrei bewegt. Auf welchen dieser Abbildungen ändern sich im Laufe der dargestellten Bewegung die Kräfte, die auf die Kugel wirken?

- ☐ geradeaus



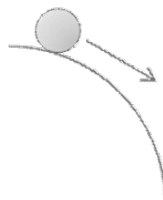
- ☐ bergab



- ☐ bergauf



- ☐ bergab





10. Ein Motorrad beschleunigt gleichmässig von 0 auf 100 km/h. Die Geschwindigkeit nimmt somit linear zu. Welche der folgenden Aussagen treffen zu?

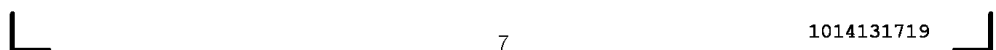
- ☐ Die Antriebskraft muss unabhängig von Luft- und Reibungswiderständen in gleichem Mass wie die Geschwindigkeit zunehmen.
- ☐ Die Antriebskraft bleibt während des ganzen Beschleunigungsvorgangs konstant, vorausgesetzt, dass sich Luft- und Reibungswiderstände nicht ändern.
- ☐ Die Antriebskraft ist zu Beginn am grössten. Sie kann langsam reduziert werden, da das Motorrad zunehmend den Schwung zur Beschleunigung ausnutzen kann.
- ☐ Wenn der Luftwiderstand mit zunehmender Geschwindigkeit wächst, dann muss auch die Antriebskraft entsprechend zunehmen.

11. Die folgenden drei Kugeln bewegen sich auf einer waagrechten Ebene:

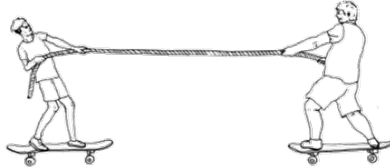
- Kugel A rollt mit der Geschwindigkeit 1 m/s um eine Kurve.
- Kugel B beginnt mit einer Geschwindigkeit von 6 m/s und wird dann immer langsamer.
- Kugel C bewegt sich immer schneller.

Welche der folgenden Aussagen treffen zu?

- ☐ Kugel A erfährt eine horizontale Kraft.
- ☐ Kugel B erfährt eine horizontale Kraft.
- ☐ Kugel C erfährt eine horizontale Kraft.



12. a) Zwei Skateboard-Fahrer mit deutlich unterschiedlichem Gewicht stehen sich je auf einem Skateboard gegenüber und sind mit einem gespannten Seil verbunden. Der linke und leichtere Skater zieht aktiv am Seil, der schwerere rechte Skater hält es nur fest. Was trifft zu?



- ☐ Sie treffen sich in einem Punkt, der näher bei der Ausgangsposition des leichteren Skaters liegt.
- ☐ Es passiert nichts, da die Kraft des Zuges eine ebenso grosse Gegenkraft hervorruft und sich die beiden Kräfte somit aufheben.
- ☐ Der leichtere Skater bleibt stehen, der schwerere Skater rollt auf ihn zu.
- ☐ Beide bewegen sich gleich schnell zur Mitte hin.
- ☐ Sie treffen sich in einem Punkt, der näher bei der Ausgangsposition des schwereren Skaters liegt.

b) Welche der folgenden Erklärungen für deine Antwort(en) ist richtig? **Bitte kreuze nur eine Antwort an.**

- ☐ Da der linke Skater am rechten zieht und nicht umgekehrt, bewegt sich der rechte Skater.
- ☐ Weil der rechte Skater das Seil ebenfalls festhalten muss, übt das Seil auch einen geringeren Zug auf den linken Skater aus.
- ☐ Der rechte Skater muss das Seil genauso fest halten, wie der linke Skater am Seil zieht. Auf beide wirkt deshalb eine gleich grosse Kraft.
- ☐ Auf den linken Skater wirkt seine eigene Kraft plus diejenige, mit der der rechte das Seil hält. Deshalb muss sich der linke Skater schneller bewegen als der rechte.
- ☐ Die Zugkraft des linken Skaters wird über das Seil zur Hälfte auf den linken, zur Hälfte auf den rechten Skater aufgeteilt.

plus

Die Erde zieht den Mond an, sonst könnte er sich nicht um die Erde bewegen. Die Masse der Erde ist 81-mal grösser als die Masse des Mondes. Welche Aussagen sind korrekt?

- ☐ Die Anziehungskraft der Erde auf den Mond ist 81-mal grösser als die Anziehungskraft des Mondes auf die Erde, weil sich die Anziehungskraft in Abhängigkeit der Masse verändert.
- ☐ Die Anziehungskraft der Erde auf den Mond und die Anziehungskraft des Mondes auf die Erde sind gleich gross, obwohl die Massen von Erde und Mond sehr unterschiedlich sind.
- ☐ Der Mond zieht die Erde gar nicht an, weil der Mond um die Erde kreist und nicht umgekehrt die Erde um den Mond.
- ☐ Die Anziehungskraft der Erde auf den Mond ist grösser als die Anziehungskraft des Mondes auf die Erde, aber nicht genau 81-mal.

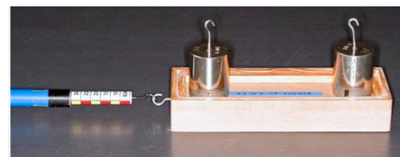
Ein Magnet (der Masse 100 Gramm) wird in die Nähe eines Eisennagels (3 Gramm) gehalten. Was trifft zu?

- ☐ Der Magnet zieht den Eisennagel an. Der Eisennagel zieht den Magneten nicht an, da der Eisennagel kein Magnet ist.
- ☐ Beide ziehen sich gegenseitig an. Die Kraft des Magneten auf den Eisennagel ist grösser, weil die Kraftwirkung vom Magneten ausgeht.
- ☐ Beide ziehen sich gegenseitig an. Die Kraft des Magneten auf den Eisennagel ist grösser, weil dieser deutlich schwerer ist als der Eisennagel.
- ☐ Beide ziehen sich gegenseitig gleich stark an.

Welche Aussagen sind bei einer Liftfahrt zutreffend?

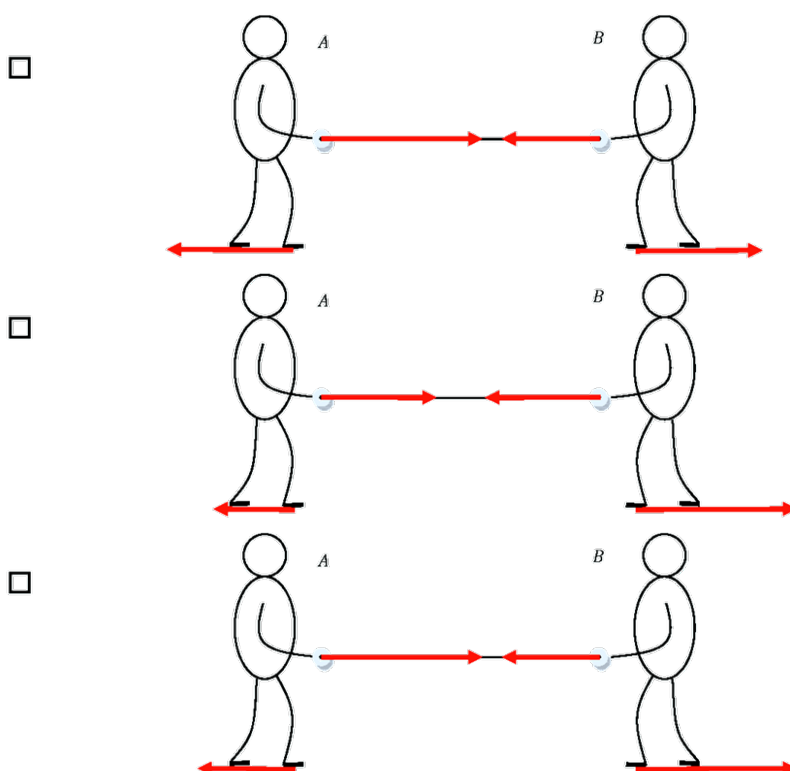
- ☐ Während der Lift mit konstanter Geschwindigkeit nach oben fährt, ist die Zugkraft grösser als die Anziehungskraft der Erde.
- ☐ Während der Lift mit konstanter Geschwindigkeit nach unten fährt, ist die Anziehungskraft der Erde grösser als die Zugkraft.
- ☐ Wenn der Lift steht, wirkt nur noch die Anziehungskraft der Erde auf den Lift.
- ☐ Beim Start vom 2. in den 4. Stock braucht es mehr Zugkraft als beim Start vom 4. in den 2. Stock.
- ☐ Wenn der Lift vom 2. Stock im 4. Stock ankommt, braucht es weniger Zugkraft beim Abbremsen, als wenn der Lift vom 4. Stock im 2. Stock ankommt.

Die Lehrperson nimmt den Kraftmesser in die Hand und zieht damit einen Holzklotz mit konstanter Geschwindigkeit über eine Tischplatte. Was zeigt der Kraftmesser an?



- ☐ Der Kraftmesser zeigt bei konstanter Geschwindigkeit keine Kraft an.
- ☐ Der Kraftmesser zeigt die Kraft an, mit der er selbst am Holzklotz zieht.
- ☐ Der Kraftmesser zeigt die Kraft an, mit der der Holzklotz am Kraftmesser zieht.
- ☐ Die Kraft, die der Kraftmesser anzeigt, ist so gross wie die Reibungskraft zwischen Tisch und Holzklotz.

A und B veranstalten ein Seilziehen. Die Kräfte sind durch Pfeile dargestellt, die jeweils an den Personen angreifen. In welcher Skizze sind die Kraftpfeile im richtigen Grössenverhältnis zueinander eingezeichnet, gegeben Teilnehmer B gewinnt das Seilziehen?



Welche der folgenden Aussagen beschreiben die Trägheit von Objekten?

- ☐ Unter Einwirkung derselben Kraft erfährt ein leichtes Objekt eine grössere Beschleunigung als ein schweres Objekt.
- ☐ Wirkt auf ein Objekt keine Kraft, so erfährt es keine Beschleunigung.
- ☐ Ein Objekt möchte aufgrund seiner Masse seinen Bewegungszustand auch bei Krafteinwirkung zunächst beibehalten.
- ☐ Wie sich eine Kraft auf den Betrag der Beschleunigung eines Objekts auswirkt, hängt von der Masse des Objekts ab.
- ☐ Reibungskräfte verzögern die Bewegung von Objekten.

Table Appendix B

List of Correct Answer Alternatives of All Items of the bMCT (plus)

Item	Answer alternatives correct
1. Water Glass	4
2. Book	2
3. Bus	1
4. Train	a3, b4
5. Walker	2
6. Wagon	1, 4
7. Object Motion	4, 5
8. Stone	3, 4
9. Inclined Plane	4
10. Motorcycle	2, 4
11. Balls	1, 2, 3
12. Skaters	a5, b3
Plus Earth	2
Plus Magnet	4
Plus Elevator	4, 5
Plus Dynamometer	2, 3, 4
Plus Tug-of-war	2
Plus Inertia	1, 2, 4

Appendix C

Quantitative Problem Solving Test

Auf den folgenden Seiten finden Sie drei Aufgaben.

Bitte bearbeiten Sie die Aufgaben gewissenhaft.

Geben Sie hier bitte noch einmal Ihren Vornamen und Namen an.

Vorname:

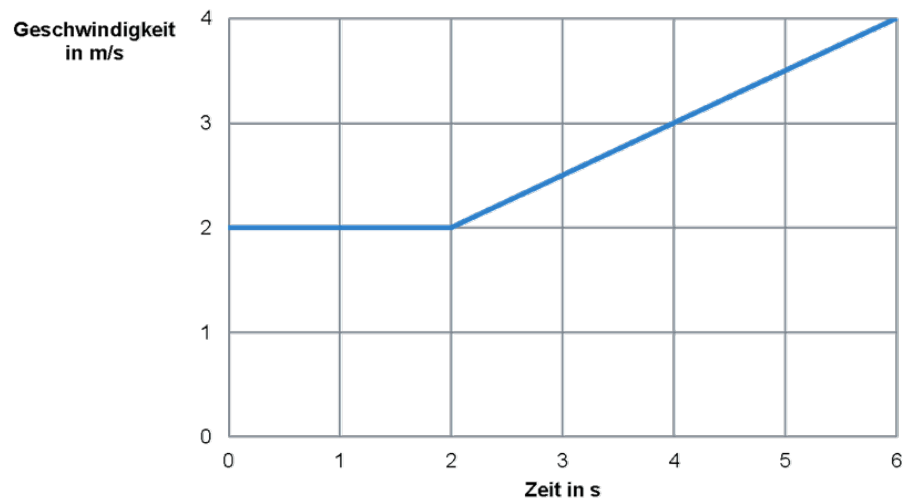
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Name:

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Vielen Dank!

28. In folgendem Diagramm ist der Verlauf der Bewegung eines Objektes (12 kg) dargestellt. Wie gross ist die resultierende Kraft, die auf das Objekt innerhalb der 6 s wirkt?



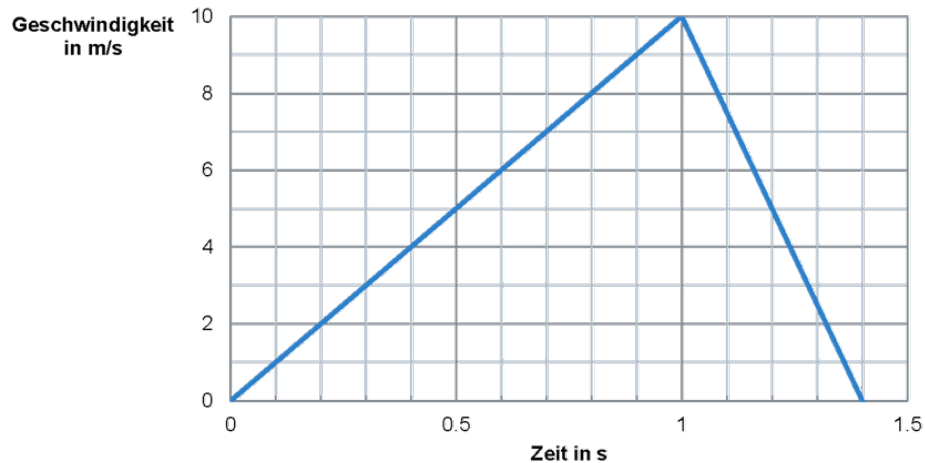
Bevor Sie zu rechnen beginnen, überlegen Sie sich bitte, welche physikalischen Begriffe und Gesetzmässigkeiten/Zusammenhänge Sie bei dieser Aufgabe berücksichtigen müssen.

Notieren Sie sich hier Ihre Ideen:

und rechnen Sie:

This is the version of the quantitative problem solving test **with scaffolds**. The text that is exemplarily highlighted on this page using the dotted box is omitted in the version without scaffolds.

29. Anna ($m = 60 \text{ kg}$) hüpft von einem Sprungbrett ins Wasser. In folgendem Diagramm ist der Verlauf ihrer Bewegung dargestellt.



- a) Berechnen Sie die Beschleunigung, die Anna in der ersten Sekunde erfährt.

Bevor Sie zu rechnen beginnen, überlegen Sie sich bitte, welche physikalischen Begriffe und Gesetzmässigkeiten/Zusammenhänge Sie bei dieser Aufgabe berücksichtigen müssen.

Notieren Sie sich hier Ihre Ideen:

und rechnen Sie:

b) Wie gross ist die resultierende Bremskraft, die auf Anna im Wasser wirkt?

Bevor Sie zu rechnen beginnen, überlegen Sie sich bitte, welche physikalischen Begriffe und Gesetzmässigkeiten/Zusammenhänge Sie bei dieser Aufgabe berücksichtigen müssen.

Notieren Sie sich hier Ihre Ideen:

und rechnen Sie:

- c) Mark ($m = 80 \text{ kg}$) springt jetzt vom Sprungbrett ins Wasser. Auch seine Bewegung folgt genau dem oben abgebildeten Diagramm. Wie gross ist die resultierende Bremskraft, die auf Mark im Wasser wirkt?

Bevor Sie zu rechnen beginnen, überlegen Sie sich bitte, welche physikalischen Begriffe und Gesetzmässigkeiten/Zusammenhänge Sie bei dieser Aufgabe berücksichtigen müssen.

Notieren Sie sich hier Ihre Ideen:

und rechnen Sie:

30. Die Trägerrakete Ariane 5 der European Space Agency (ESA) dient heute dazu Kommunikationssatelliten in eine geostationäre Umlaufbahn zu schießen.

Die Triebwerke erzeugen beim Start einen Schub von $12 \cdot 10^6 \text{ N}$. Das Gewicht der Rakete beträgt $8 \cdot 10^6 \text{ N}$.

Berechnen Sie die Startbeschleunigung der Rakete.



Bevor Sie zu rechnen beginnen, überlegen Sie sich bitte, welche physikalischen Begriffe und Gesetzmässigkeiten/Zusammenhänge Sie bei dieser Aufgabe berücksichtigen müssen.

Notieren Sie sich hier Ihre Ideen:

und rechnen Sie:

Vielen Dank!