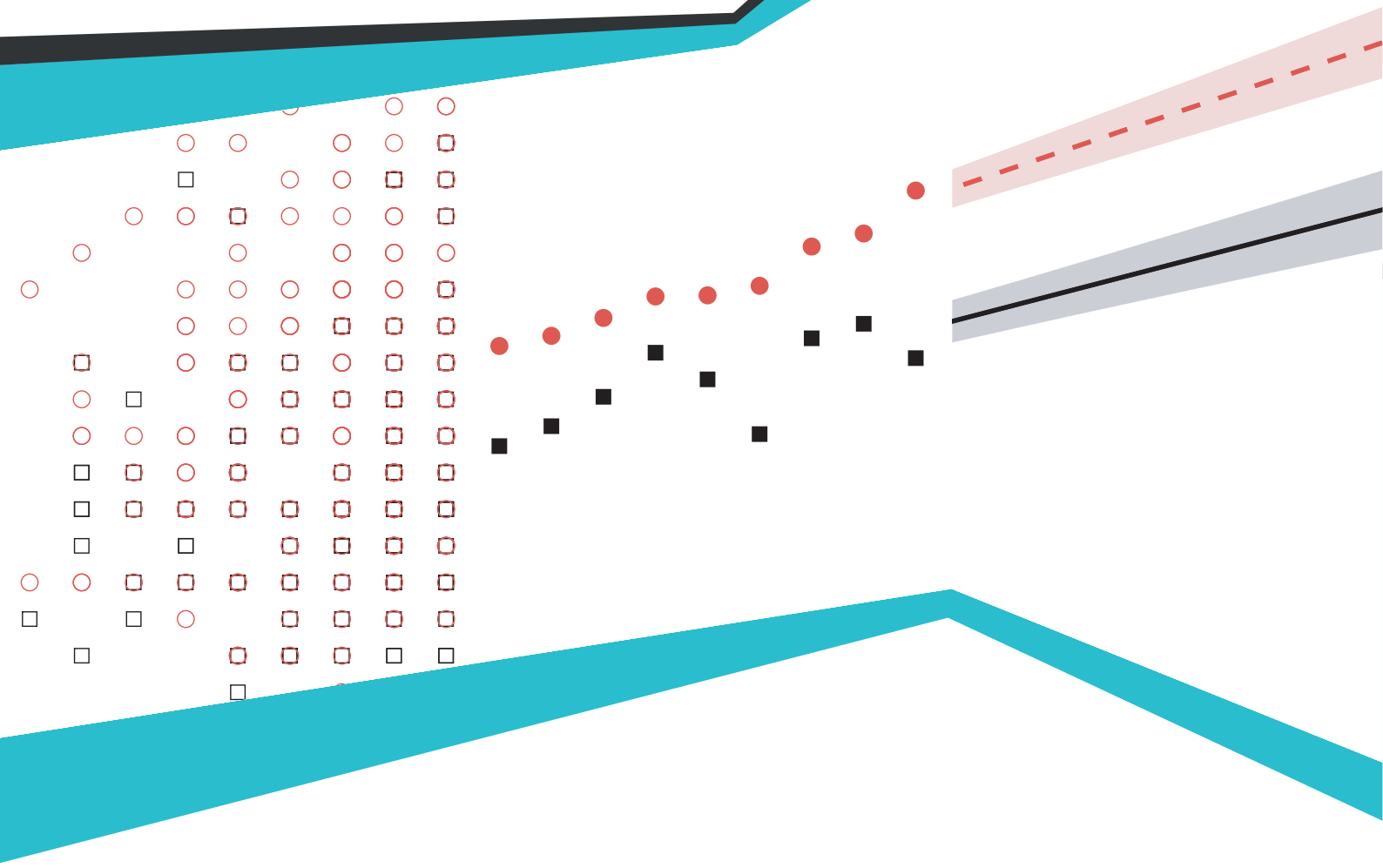
2 Hamburg Advances in Science and Engineering Education Research
Editor: Christian Kautz



Assessing the Effectiveness of Research-Based Active Learning Materials for Introductory Engineering Mechanics

Julie Direnga

ENGINEERING EDUCATION RESEARCH
FACHDIDAKTIK DER INGENIEUR-WISSENSCHAFTEN
TUHH
Hamburg University of Technology

Hamburg Advances in Science and Engineering Education Research
Editor: Christian Kautz

Volume 2

ASSESSING THE EFFECTIVENESS OF RESEARCH-BASED ACTIVE LEARNING MATERIALS FOR INTRODUCTORY ENGINEERING MECHANICS

by Julie Direnga

DOI: https://doi.org/10.15480/882.3229

Hamburg Advances in Science and Engineering Education Research, Volume 2
Editor: Christian Kautz

Imprint

Engineering Education Research Group
Hamburg University of Technology
Am Schwarzenberg-Campus 3
21073 Hamburg
Germany

License

1st edition, April 2021

DOI: https://doi.org/10.15480/882.3229
Julie Direnga: https://orcid.org/0000-0002-3936-7032

ASSESSING THE EFFECTIVENESS OF RESEARCH-BASED ACTIVE LEARNING MATERIALS FOR INTRODUCTORY ENGINEERING MECHANICS

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktor-Ingenieurin (Dr.-Ing.)

genehmigte Dissertation

von
Julie Direnga

aus
Bremen

2021

1. Gutachter: Prof. Dr. Christian Kautz
Abteilung für Fachdidaktik der Ingenieurwissenschaften
Technische Universität Hamburg

2. Gutachter: Prof. Dr.-Ing. Andreas Baumgart
Department Maschinenbau und Produktion
Hochschule für Angewandte Wissenschaften Hamburg

3. Gutachter: Prof. Shane Brown, Ph. D.
Civil & Construction Engineering
Oregon State University

Datum der mündlichen Prüfung: 23. Oktober 2020

ACKNOWLEDGMENTS

Although a dissertation is authored by one person alone, it "takes a village" to complete it. Along the way - and this includes times at which the thought of a Ph.D. had not even occurred to me - I have been supported, guided, and inspired by many people who should not be left unmentioned.

First and foremost, I would like to thank Prof. Dr. Christian Kautz, head of the *Engineering Education Research Group (FD)*, for his inspirational teaching during my freshman year - which sparked my interest in this kind of research in the first place - and for later introducing me to the fascinating research and the fast-growing community of engineering education. As my advisor, he showed me new ways of thinking, honest appreciation for my work, and provided guidance when needed. Furthermore, I would like to thank Prof. Dr.-Ing. Andreas Baumgart and Prof. Shane Brown, Ph.D. for reviewing my dissertation, and Prof. Dr. Norbert Hoffmann for acting as the chairman of the examination board.

I am also deeply grateful to my colleague and friend Dr. Dion Timmermann for paving the way. He contributed substantially to giving me a smooth start in the research group and (unless "Poldi" was guarding his door) he always was available as an invaluable "sparring partner" on sometimes difficult intellectual terrain. Furthermore, Dion's continuous dedication to bringing the research group forward by critically reflecting on our work and our own mental models of our science shaped the environment in which both our dissertations could grow and finally flourish.

In the current head and founding member of the *Center for Teaching and Learning (ZLL)*, Dr. Andrea Brose, I found another such environment shaper and friend who always puts a smile on my face (and anything on pizza). Her early work in the FD together with Christian Kautz laid the foundation for my research and it was an honor for me to continue on this path.

I thank *all* of the ZLL/FD team members, who created a fruitful and warm atmosphere that was one of a kind. My special thanks among the team members go to my late office partner and friend Ferdinand Kieckhäfer, who greatly supported me in focusing my research topic. Plus, whenever I ran low on motivation, a "Kofi" and countless "Promodoro" sessions helped me get back on track.

Education research - whether discipline-based or not - relies on people for data. I hereby thank all students who completed a test or volunteered for an interview, all lecturers who granted access to their

students for such tests, as well as all experts who spared time for me to formally interview them.

Informally, I sometimes had to reach out to CATS-creator Prof. Paul Steif, Ph.D., whom I thank for always readily answering my questions about CATS history and for the permission to print the items.

Last but not least, I want to thank my family, especially my wife Karin and all my parents for the love and support they have given me before, during and after completion of this work.

ABSTRACT

Functional understanding of fundamental mechanics concepts is essential for successful problem-solving of complex engineering tasks. Especially introductory mechanics courses as the basis for subsequent instruction must ensure that students reach a level of conceptual understanding that enables them to make sense of higher-level content. Research shows that traditional instruction may not achieve these goals. Instead, research-based instruction is often more successful.

This dissertation investigates concept inventory data from pre- and post-tests administered in an introductory Statics course to assess the effectiveness of research-based active-learning materials (RBALM) in the form of *Tutorials* (Kautz et al., 2018) in fostering student conceptual understanding. The *Tutorials* were designed after the *Tutorials in Introductory Physics* (McDermott and Shaffer, 1998), a collection of collaborative-group worksheets that do not involve extensive calculations but focus on understanding the concepts. For effective discussions, the materials are intended to be used in multiple small groups of three to four students under the supervision of teaching assistants who have been trained to engage the students in Socratic dialogue.

The dissertation is structured in three parts. In Part I, the instrument administered as post-test, the Concept Assessment Tool for Statics (CATS), is closely examined for validity as a measure for student conceptual understanding in the given context of a German higher education introductory mechanics course. While it is found that the CATS may be used as a post-test, validity issues when using it with students prior to Statics instruction justifies and necessitates the use of a different, more suitable instrument as pre-test, i. e. the Force Concept Inventory (FCI). The challenge of interpreting data collected with non-identical pre- and post-tests (NIPPs) motivates the development of the Discriminative Learning Gain (DLG), a two-parameter quantification of the difference in learning success between courses, which is introduced and discussed in Part II. Using this analysis method in Part III, FCI pre- and CATS post-test data from 10 cohorts, of which some used *Tutorials*, are compared. In addition, a longitudinal study is conducted to investigate the long-term effects of the *Tutorials*. The results of the DLG analysis allow to conclude that students of all pre-test levels tend to benefit from the RBALM, and especially students with high pre-test scores. The longitudinal study reveals that students who were taught without *Tutorials* often only reach a similarly high level of understanding as the students with *Tutorials* when they had engaged in teaching assistant activities after their own instruction.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

1PL	one-parameter logistic 89
2PL	two-parameter logistic 64
3PL	three-parameter logistic 89
ABET	Accreditation Board for Engineering and Technology . . 10
AIC	Akaike information criterion 159
ANCOVA	analysis of covariance 187
ANOVA	analysis of variance 61
ASCI	Alternative Statics Concept Inventory 48
CATS	Concept Assessment Tool for Statics 5
CI	concept inventory 4
CFA	confirmatory factor analysis 61
CTT	Classical Test Theory 42
DCM	diagnostic classification modeling 63
DBER	discipline-based education research 19
DLG	Discriminative Learning Gain 6
EER	engineering education research 20
EFA	exploratory factor analysis 63
FBD	free-body diagram 15
FCI	Force Concept Inventory 4
FMCE	Force and Motion Conceptual Evaluation 24
ICC	item characteristic curve 88
IPPs	identical pre- and post-tests 184
IRC	item response curve
IRT	Item Response Theory 42
JiTT	Just-in-Time-Teaching 214
JiTTl	JiTT "light"
MBT	Mechanics Baseline Test 45
MCQ	multiple-choice question 40
MoM	Mechanics of Materials 213
NIPPs	non-identical pre- and post-tests 5
PCA	principal component analysis 94
PEG	Physics Education Group 25

PER physics education research 19
RBALM research-based active-learning materials 1
RI retention interval 220
S Statics of rigid bodies 213
SCI Statics Concept Inventory 46
SOMCI Strength of Materials Concept Inventory 49
STEM science, technology, engineering, and mathematics 2
TA teaching assistant 25
TUHH Hamburg University of Technology 26
WLR weighted linear regression 198

1

INTRODUCTION

This dissertation is concerned with measuring and assessing the effectiveness of research-based active-learning materials (RBALM). It will demonstrate that such materials foster student conceptual understanding in introductory engineering mechanics to a greater extent than traditional instruction. In the following paragraphs, this chapter motivates three research questions which eventually lead to this and further conclusions.

The relevance of engineering mechanics is highlighted by Meriam and Kraige in the preface to their textbook on this subject:

> *"Engineering mechanics is both a foundation and a framework for most of the branches of engineering. . . . Thus, the engineering mechanics sequence is critical to the engineering curriculum."*
>
> (Meriam and Kraige, 2008, p. vii)

Taking this statement one step further, Statics, as the first course in the engineering mechanics sequence, must be especially critical to the engineering curriculum. As an introductory course, Statics serves as the basis for the subsequent engineering mechanics courses (Steif, 2004). Not only the problem-solving techniques, but also the concepts introduced in a Statics course, such as equilibrium, external or internal forces and moments, supports, or the center of mass, remain relevant in advanced courses such as Mechanics of Materials, Dynamics, Design, or Vibration Theory.

While the content of these courses evolves rather little over time, the challenges which future engineers must face tomorrow are yet unknown. Engineering education must not only prepare students to apply their knowledge in familiar situations, but also to transfer it to new problems, to identify connections between pieces of knowledge, and to organize their own life-long learning. Factual and procedural knowledge, which is required to systematically approach and solve standard problems, are not sufficient for this task. In addition, the underlying fundamental concepts must be thoroughly understood.

Conceptual understanding does not compete against factual or procedural knowledge. On the contrary, it helps to understand the logic behind a process, which in turn facilitates remembering the steps and making less errors. It also enables to extract relevant information more easily in unfamiliar situations and to recognize patterns. As only "[d]eep understanding of subject matter transforms factual information into usable knowledge" (Bransford et al., 1999, p. 16), it

must be treated as a central learning objective, which must be addressed by appropriate and effective learning materials.

Not all instruction methods foster conceptual or deep understanding to the same extent. Education research has brought forward an overwhelming amount of evidence that the teacher-centered approach, the so-called "sage on the stage", is not very effective in helping students overcome conceptual difficulties (Heron, 2015). Instead, many studies on various active-learning methods suggest a positive effect of active instructional formats on understanding of science, technology, engineering, and mathematics (STEM) concepts (Van Heuvelen, 1991; Hake, 1998; Bransford et al., 1999; McDermott, 2001; Deslauriers et al., 2011; Freeman et al., 2014; Deslauriers et al., 2019). This phenomenon can be explained by the framework of constructivist learning theories (e. g. Piaget, 1970; diSessa, 1993; Vosniadou, 1994), which assume that knowledge cannot be transferred from teacher to learner but must be actively constructed by the individual. In this process, prior experiences and conceptions serve as anchors and fundaments to newly constructed knowledge. These preconceptions are often *misconceptions* in the sense that they do not align with the conceptions of experts in the field, and thus must be changed to achieve expert understanding.

Conceptual change theory suggests that a shift from a naïve preconception to the expert conception is unlikely to be achieved by student activity alone. The success of such a change of conceptions depends on many factors, such as the type of misconception (Chi, 2013), or the level of dissatisfaction the student experiences with their current conception (Posner et al., 1982). Evidence for this perspective has been published for instance by Andrews et al. (2011) and Deslauriers et al. (2011). To foster conceptual change, student preconceptions must thus be adequately addressed by instruction.

RBALM are developed based on evidence about student preconceptions (e. g. Heron et al., 2003; McBride et al., 2010). They address difficult concepts in an activating, student-centered setting. In the framework of constructivism and conceptual change theory, RBALM are likely to be more effective in promoting conceptual understanding than instruction that does not consider such research results or that is teacher-centered, but in order to ensure that students indeed benefit from such materials, their effectiveness must be scientifically assessed.

One example of RBALM are the *Tutorials in Introductory Physics* by McDermott and Shaffer (1998) and their implementation in physics education at various institutions. These are worksheets designed to address known misconceptions and thereby to foster conceptual change. They were developed based on qualitative and quantitative research results obtained through student interviews and written questions on various concepts. The *Tutorials* are meant to be imple-

mented in small-group discussions, occasionally aided by simple experiments, under the supervision of an instructor trained in Socratic dialogue. They are not only research-based in their development but also evidence-based with respect to assessment of their effectiveness, for which they have been described as "one of the most widely researched physics education research-(PER-)based reforms" (Finkelstein and Pollock, 2005, p. 1).

While there are substantial overlaps in fundamental concepts taught in introductory physics and engineering instruction, engineering mechanics is different from physics. The goal of physics is to discover physical principles and how they can explain observable phenomena. In Physics, the inspected systems are often simplified as point masses such that the aspect of moments and rotation is avoided (Steif, 2004). These very simple models of a system are often sufficient because the focus of interest lies on the physical principles and not on the system (Gross et al., 2011, p. 3). For example, early physics instruction often teaches moment equilibrium separately from and considerably later than force equilibrium. First, all principles are examined in case of linear motion, then, the same is repeated for rotation. In Statics, the concepts of force equilibrium (linear) and moment equilibrium (rotation) have to be considered simultaneously (Newcomer and Steif, 2008). The goal of engineering mechanics is to describe and predict forces or motion in real-world systems, using the (rather few) underlying physical principles as tools. The inspected systems may become arbitrarily complex, consisting of multiple extended bodies, connected by multiple types of joints, and subjected to multiple or distributed forces. In order to be able to keep problem-solving of such complex systems manageable, engineering mechanics emphasizes a very systematic problem-solving approach (e. g. Brommundt et al., 2007, p. 41), which results generally in a rather deductive way of learning the concepts. In contrast, the goal of Physics education rather suggests an inductive way of learning (although also Physics instructors often choose deductive over inductive reasoning (McDermott, 1991, p. 304)).

Furthermore, for engineers and their field of work, "the static state is an important special case of motion" (Gross et al., 2011, p. 1, translated)[1], which justifies spending an entire semester on it. Physicists treat statics as a (not so special) case of dynamics, where the only interesting difference lies in the concept of static versus kinetic friction. Steif (2004) warns against downplaying the difficulties that may arise in learning Statics:

> *"The only scientific principle in Statics, the principle of equilibrium, which is captured by the net force equaling zero, is merely a subset of what is taught in physics. Surely, Statics must be a breeze for students who have passed physics! This deceptive simplicity of*

1 German original: "Ein wichtiger Sonderfall der Bewegung ist der Zustand der Ruhe."

> *Statics can lead, unfortunately, to instruction that is insufficiently sensitive to the subtleties of implementing the equilibrium principle [...].”*

All these differences require instructional materials specially designed for engineering mechanics. Kautz et al. (2018) adopted the approach advocated by McDermott (2001) to design *Tutorials* for engineering mechanics. Following the same research-based philosophy, they conducted systematic investigations to first identify the conceptual difficulties that students encounter in engineering mechanics (Brose and Kautz, 2011). The results were then used to develop collaborative-group worksheets that address these difficulties as potential misconceptions. In this dissertation, the effectiveness of RBALM with respect to strengthening student conceptual understanding in introductory mechanics courses will be investigated using parts of these materials. The investigation is guided by the following question:

RESEARCH QUESTION 1 Is instruction that uses research-based active-learning materials (RBALM) in the form of *Tutorials* more effective in promoting student conceptual understanding than traditional instruction in the context of introductory engineering mechanics courses?

Before this question can be answered, however, methodological questions of how to assess the effectiveness of instruction must be addressed. Assessment of learning material effectiveness has to be performed with adequate measurement tools and analysis methods. Many studies still rely on students’ self-reported feeling of learning, although it has now been shown to be a highly inadequate and even misleading measure (Deslauriers et al., 2019). Instead, STEM education researchers often use standardized instruments as pre- and post-tests to assess the effect of a course or intervention on student learning (e. g. Hake, 1998; Beichner, 2009; Pollock, 2009; Andrews et al., 2011; Ding and Liu, 2012; Moskal et al., 2013; Theobald and Freeman, 2014). With this method, the two tests are performed right before and after the period of instruction. The pre-test thereby provides a baseline measurement which becomes essential if one wants to compare populations with possibly different levels of prior knowledge in non-randomized studies. The post-test gives insight on the absolute level of understanding at the end of the period of interest.

The development and application of standardized test instruments such as concept inventories (CIs) have fostered the discussion on and innovation of teaching and learning. The Force Concept Inventory (FCI), developed in 1992 by US physics instructors Hestenes et al. (1992) addresses the Newtonian force concept in 30 multiple-choice questions that require no calculations. Many of their colleagues were surprised when their students failed on the supposedly easy test questions. Realizing that this lack of understanding had remained largely

undiscovered in the frame of traditional instruction sparked a number of innovations (e. g. Mazur, 1997; Hestenes and Wells, 1992). But the FCI did not only serve as a spark, it continued to serve as a standardized assessment tool to compare results over various courses and implementations of instruction (e. g. Hake, 1998). The success of the FCI has led to the development of CIs for other concepts and other disciplines, predominantly in the US. As of today, more than 150 grants have been awarded by the National Science Foundation to projects mentioning the terms "concept inventory" or "concept inventories" in their project titles or abstracts (National Science Foundation, 2018). If a measurement tool on a specific topic is required, existing CIs should thus be preferred instead of "reinventing the wheel".

Yet, often, re-establishing validity of the test results is necessary, for example, if the instrument is adapted to a different national context or when the population changes (Lindell and Ding, 2013). According to Crocker and Algina (1986), validation is "the process by which a test developer or test user collects evidence to support the types of inferences that are to be drawn from test scores". Apart from possible translation issues in a different national context, incompatibilities in terms of notations, conventions or course content may affect the validity of the test result interpretation. Even for identical course content, different learning goals might be pursued in different geographical regions. Therefore, results of validity studies do not necessarily apply when the national context differs. In such a case, the validity study should be repeated. The central measurement instrument used in this dissertation, the Concept Assessment Tool for Statics (CATS), was developed in a different national context which motivates the second research question:

RESEARCH QUESTION 2 Is the Concept Assessment Tool for Statics (CATS) a suitable measurement instrument for the investigated context of an introductory engineering mechanics course in the German higher education system?

The same issue arises when the instrument is administered to a different population. Evidently, a test developed for university level courses may not be suitable for testing elementary school children. Likewise, an instrument may not be suitable as both pre- and post-test, because instruction changes the population. If no difference in pre-instruction knowledge is expected, pre-testing might not be necessary, for example, when it is very unlikely that students have any prior knowledge on the topic in question. In case a baseline measurement is required and there is no suitable instrument for both populations (pre- and post-instruction), one option is the use of two instruments as non-identical pre- and post-tests (NIPPs). While there are established statistical measures for reporting results from identical pre-and post-tests like the average normalized gain (Hake, 1998)

or normalized change (Marx and Cummings, 2007), an equivalent measure for quantifying the "learning gain" measured with NIPPs is missing. In this dissertation, the *Discriminative Learning Gain (DLG)* is proposed as such a measure as an outcome of the third research question:

> RESEARCH QUESTION 3 How can data collected with non-identical pre- and post-tests (NIPPs) be interpreted and processed to quantify any effect of the materials?

Following the research questions (in a slightly different order), the dissertation addresses three main investigations:

1. In Part I, the CATS, an instrument developed in the US for measuring conceptual understanding of statics is analyzed for validity in the German higher education context,
2. in Part II, the DLG, a statistical analysis method tailored to the specifications of data from NIPPs is proposed, and
3. in Part III, the effect of RBALM, instructor expertise, and other variables on student conceptual understanding in a German introductory engineering mechanics course are investigated.

Each of these parts ends with a conclusion of its own. In the final conclusion, results from all parts are considered and further higher-order questions are addressed.

Engineering education research uses theory and concepts from both parent disciplines, engineering and education. Most of the research originates from the US, while this dissertation is set in the German context. As not all readers may be familiar with both those disciplines and both contexts, the basics required to understand all parts are briefly introduced in the following background chapters. It is suggested to read them selectively as needed.

BACKGROUND

2

THE GERMAN HIGHER EDUCATION SYSTEM

This research is conducted in the context of the German higher education system with focus on engineering mechanics in a mixed-methods approach. Because much of the research done on engineering education has been conducted in the US context, the specifications of the German higher education and course context and the aspects in which they differ from the US context shall be briefly addressed here.

FIRST-YEAR CURRICULUM US engineering study programs rarely offer Statics courses in the first year of study. Instead, the students enrolled in Statics courses are generally sophomores (second-year students) or juniors (third-year students) (Steif and Hansen, 2007). First-year requirements often involve Mathematics, Physics, and sometimes introductory courses on engineering (Steif and Hansen, 2006a) that aim at giving an overview of the discipline and sometimes involve labs or small design projects (e. g. Cornell University). In contrast, at German universities, Statics is mostly taught in the first semester (e. g. Hamburg University of Technology, Technical University of Munich, Karlsruhe Institute of Technology, Technische Universität Berlin RWTH Aachen University). An introductory course on engineering or Physics is offered as a co-requisite in some institutions (e. g. RWTH Aachen University and Technical University of Munich, respectively), but the importance that is attributed to such courses differs among the institutions.

FOCUS ON COURSES VS FOCUS ON EXAMS In the US system, it is common for students to register for a course, which includes one or more exams (final and midterms). Graded homework assignments are often part of the course and attendance can be required to receive a satisfactory grade. Often, there are several parallel sessions, taught by different instructors who must agree upon identical content. Radical changes to the course must be granted by a board at the departmental level.

German universities have been shaped by the Humboldtian model of higher education in various aspects. For instance, following the principle of academic freedom, students are attributed the status of autonomous learners who are free to chose when, how, and what to learn. The regulations have become stricter over time and there are exceptions, but the basic idea remains prominent. By the same principle, instructors are free to teach whatever content with whatever methods they choose. In case of radical changes to a course, colleagues may in-

formally criticize the instructional design and implementation, but this criticism is often the metaphorical "tiger without teeth". Yet, the freedom of the instructor is limited by the freedom of the students, so certain instructional designs may create a conflict. For example, it is difficult for instructors to implement graded course-work or graded in-class activities as student attendance in class is not required. Students do not register for a course, they register for the exam to that course. Students who chose to attend the course but then failed the exam must resit the exam but are not required to attend the course a second time. How they acquire the necessary skills and knowledge to pass the exam is their free choice. Consequently, the exams must be comparable each year, which may hinder changes to instruction if constructive alignment is taken seriously in that the defined learning outcomes are appropriately reflected by the exam and addressed accordingly by instruction.

INTENDED LEARNING OUTCOMES The focus on intended learning outcomes was set through the Bologna Process, which was initiated in 1988 to strengthen employability and student mobility within Europe, and which softened these principles of freedom. One of its consequences was the "Kompetenzorientierung", setting the focus on the competencies students should have after finishing their program. At that time, discussions on an outcome-based approach to accrediting study programs had already started in the US, which finally led into the adoption of the Engineering Criteria 2000 (EC2000) by the Accreditation Board for Engineering and Technology (ABET) in 1997 (ABET). Despite this similar development, the acceptance and implementation of outcome-based instruction design is much more widespread and made explicit in the US than in Germany.

COMMON FORMATS OF ASSESSMENTS Another difference, which is relevant to this research shows in the use and acceptance of assessment with multiple-choice-instruments, which is widespread in the entire US education system, unlike in Germany, where open-ended exam questions are the norm.

These differences should be kept in mind when interpreting and comparing results and implications for research and instructional practice in the different contexts.

3

REVIEW OF STATICS CONCEPTS

In this chapter, the fundamental concepts introduced in the introductory engineering mechanics course on statics are reviewed. Readers who are familiar with these concepts are encouraged to skip this chapter with the option to consult it later, if necessary. This text is not meant to replace a textbook on statics. For more details and examples, please refer to a statics textbook of your choice.

3.1 INTRODUCTION

Mechanics is the oldest branch of physics (Gross et al., 2011; Meriam and Kraige, 2008). Its task is to describe and predict the motion of bodies or parts of bodies and the forces acting between them. It can be categorized with respect to different aspects. Three examples are named by Gross et al. (2011) and illustrated in Figure 1.

The first (a) focuses on the state of aggregation of the inspected bodies. These behave differently whether they are in a solid, liquid or gaseous state. Furthermore, solid bodies may be regarded as rigid if their deformation under a load is small. In case the deformation is of interest, different models for elastic or plastic deformation may apply.

The second example of internal structure (b) discriminates between kinematics and dynamics. The former is concerned with describing only the motion of bodies while the latter makes statements about the forces causing or resulting from this motion. The type of motion can then be categorized into accelerated (kinetics) and unaccelerated (statics). In contrast to physics education, where statics is only viewed as a special case of dynamics in which the bodies are not accelerated, statics is of special importance in engineering education because many applications are designed for the static case. Often, the introductory mechanics course only considers rigid bodies in static equilibrium (*Statics*). The subsequent course then covers elastically-deformable bodies in static equilibrium (*Mechanics of Materials*). After that, kinetics are addressed. This course is usually referred to as *(Engineering) Dynamics*. The nomenclature of the courses is thus slightly different than the categorization presented by Gross et al. (2011).

The third example (c) categorizes Mechanics into analytical mechanics and engineering mechanics. While analytical mechanics is part of the physics discipline and thus concerned with discovering basic principles and laws of nature, engineering mechanics focuses on helping engineers solve specific problems.

The part of mechanics addressed in this study is engineering mechanics of rigid solid bodies in static equilibrium.

The field *statics of rigid bodies* is based on axioms, which are basic assumptions underlying a theory. They cannot be proven, but they reflect the outcome of empirical experiments accurately enough to be accepted as true without proof. Neither the number nor the sequence of these basic assumptions is unique in the literature (Dankert and Dankert, 2013, p. 3). Some textbook authors address only a minimum of four axioms which are truly independent (e. g. Dankert and Dankert, 2013; Eller, 2015), others list as many as ten "out of convenience" (Brommundt et al., 2007, p. 11), and some do not explicitly list them at all (e. g. Gross et al., 2011). To show the diversity, Figure 2 illustrates how three different textbooks present the axioms differently. Among those, the two textbooks that introduce only four axioms do not even fully agree.

No real body is truly rigid but will always deform under an applied load, even if this deformation may be undetectably small, or the applied load must be larger than reasonable for a measurable effect. Some axioms are not affected by the idealization of assuming the bodies as rigid. For example, Newton's Third Law, which is addressed by **Action = reaction**, always applies, even for deformable bodies. For other axioms, the idealization of rigid bodies is essential. For example, for deformable bodies, forces acting at different points along their lines of action do have different effects on the body, as different parts are deformed.

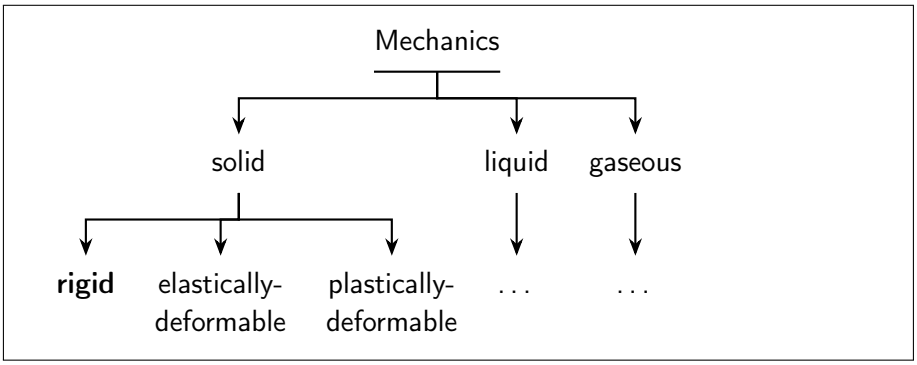
3.2 CONCEPTS ADDRESSED BY THIS DISSERTATION

The following paragraphs introduce the statics concepts that are repeatedly used in this dissertation.

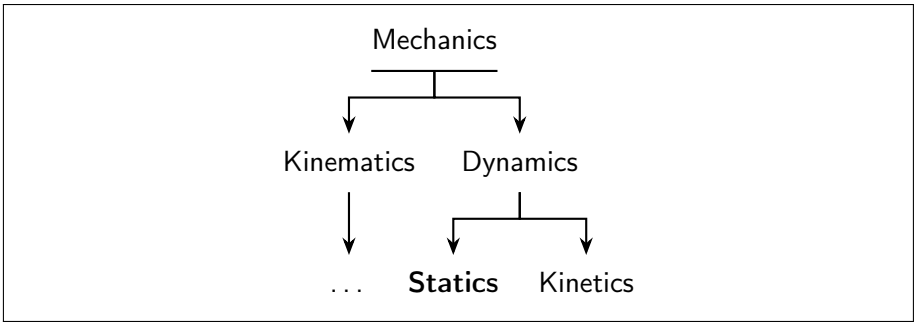
CENTER OF MASS OR CENTROID The center of mass of a body is the single point at which the body could be suspended in any orientation without experiencing angular acceleration (i. e. the sum of moments due to gravitational forces alone is zero, assuming a uniform gravitational field).

The centroid is the geometrical center of a body. For bodies of homogeneous density, the center of mass and the centroid coincide. Note that both points may lie outside of the body.

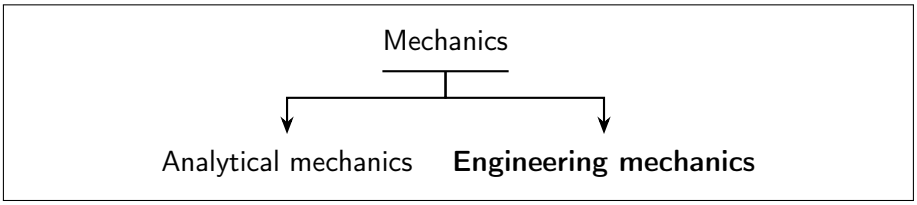
FORCES AND MOMENTS Bodies may be subjected to a variety of forces. A force is a vectorial quantity with a magnitude, a direction, and a point of action. Forces acting on the same body can thus be added. Likewise, forces can be composed into components acting in different directions of interest. If the vectorial sum of all forces is non-



(a) by state of aggregation of the bodies



(b) by the quantities of interest



(c) by discipline or character of the problem

Figure 1: Different possibilities of structuring the field of mechanics mentioned in Gross et al. (2011). Focus of the introductory course investigated in this study is indicated with bold text.

Brommundt et al. (2007)

1. **Equilibrium** Two forces are in equilibrium and cancel their effect on the rigid body if they are equal in magnitude and acting in opposite directions on the same line of action.
2. **Force vector parallelogram** Two forces acting at the same point have the same effect on a body as a single net force according to their vectorial sum.
3. **Adding or removing a set of forces in equilibrium** does not change the state of the body. (The axiom **Line of action** is presented as a consequence of this axiom.)
4. **Transferring a set of forces in equilibrium to another body** A set of forces in equilibrium has the same effect if transferred to a different free body.
5. **Free-body diagram** The equilibrium state of a body does not change if its supports are replaced by equivalent forces.
6. **Action = reaction** Any force exerted from one body on the other is opposed by an equal and opposite force.
7. **Forces between smooth bodies** When in contact, smooth bodies exert forces normal to their common tangential plane, directed towards the other body.
8. **Equilibrium of non-rigid bodies** A free, *non-rigid* body is in equilibrium if and only if it were in equilibrium as a rigid body.
9. **Systems of bodies** A system of bodies is in equilibrium if and only if all single bodies in the system are in equilibrium.
10. **Moving bodies** (part of **Newton 1**) A moving, free, rigid, and massless body is in equilibrium if and only if it is constantly in static equilibrium as seen from a reference system moving along with the body.

Dankert and Dankert (2013)

1. **Line of action** Forces can be moved along their line of action without changing their effect on the body.
2. **Force vector parallelogram**
3. **Equilibrium**
4. **Action = reaction**

Eller (2015)

1. **Newton 1** Every body remains at rest or moving at constant velocity as long as no force is acting on it.
2. **Line of action**
3. **Force vector parallelogram**
4. **Action = reaction**

Figure 2: Three versions of the axioms of rigid-body statics by three different German textbook authors.

zero, a net force acts on the body which causes linear acceleration ($\Sigma\vec{F} = m\vec{a}$).

In addition, bodies may be angularly accelerated. Any force that does not pass the body's center of mass contributes to the sum of moments about this point by $\vec{M} = \vec{F} \times \vec{d}_c$, where $\vec{d}_c$ is the position of the center of mass measured from any point on the line of action of the force. If the vectorial sum of all moments is non-zero, a net moment acts on the body which causes angular acceleration ($\Sigma\vec{M} = I\vec{\alpha}$, where I is the moment of inertia and α is the angular acceleration).

FORCE COUPLE Two forces with equal magnitudes but acting in opposite directions along the same line of action cancel their effect on the body. Both, the net force and net moment are zero. If, however, their lines of action are parallel but at a distance d between them, the effect is not cancelled. While the net force is still zero, the net moment is not. The two forces are called a force couple and they may be represented by a curved arrow ($\curvearrowleft$ / $\curvearrowright$) instead of the two forces. As for rigid bodies local deformations are irrelevant this couple may be applied anywhere on the body without changing its effect.

FORCE SYSTEMS AND EQUILIBRIUM One central task of mechanical engineering is to determine stresses due to loads. To do so, not only the geometry but also the forces at the points of interest must be known. Modeling all forces, known and unknown, acting on the body results in a system of forces which may be analyzed. The concepts of force couples and moments due to a single force are essential elements for such an analysis. In statics, all bodies are assumed to be in *static equilibrium*, i. e. they are not accelerated so that the vectorial sums of all forces and moments, respectively, must be zero ($\Sigma\vec{F} = m\vec{a} = 0$ and $\Sigma\vec{M} = I\vec{\alpha} = 0$). This condition can be used to determine unknown values.

SUPPORTS AND REACTION FORCES In order to bear loads, mechanical parts must be supported. The supports exert *reaction forces* so that equilibrium is maintained. Forces and couples can only act in those directions in which movement is restricted by the support or by contact to another body. The types of support or the characteristics of the contact thus provide information in which directions the reaction forces on the body can or cannot act. This information must be used to reduce the number of unknown forces in the system.

FREE-BODY DIAGRAMS The construction of free-body diagrams (FBDs) is "the most important single step in the solution of problems in mechanics" (Meriam and Kraige, 2008, p. 110). It involves identifying which forces act on a chosen part of the system and visualizes the interactions between parts. A system of equations can be derived

from the FBD to solve for unknown forces. The construction process itself involves the following steps (Meriam and Kraige, 2008, p.114):

1. Choosing an appropriate part of the system.
2. Isolating and reducing the part to the relevant features (e. g. by drawing only the boundary)
3. Identifying all forces acting *on* the isolated part.
4. Choosing an appropriate coordinate system.

INTERNAL FORCES AND MOMENTS The forces acting inside of a body are often of major interest to engineers. If a system is in equilibrium, so is every part of the system, even parts of bodies, which are held in place by internal forces. When isolating part of the body in a FBD, these internal forces are treated as external to the part and must be included in the FBD. By applying the equilibrium conditions and solving for the unknown internal forces, these can be determined as a function of a geometrical coordinate.

FRICTION Friction forces are contact forces that are exerted parallel to a rough (i. e. not ideally smooth) surface. The concept of friction differs between static and kinetic cases. If the surfaces move relative to each other, the friction force is determined as the product of the kinetic coefficient of friction, which depends on the surfaces, and the normal force acting between these surfaces (kinetic friction, $f = \mu N$). In the static case, friction is a reaction force, which is determined by the equilibrium conditions. It adapts to the applied loads and its magnitude can take any value between zero and the limit given by the product of the static coefficient of friction and the normal force ($f \leqslant \mu_0 N$). In case the applied forces are such that this limit is exceeded, the kinetic case will apply.

ACTION = REACTION Forces always act between bodies. The force exerted by body A on body B is *always* equal and opposite to the force exerted by body B on body A, independent of any other influences or characteristics of the bodies.

$$\vec{F}^{AB} = -\vec{F}^{BA} \tag{1}$$

This is true for all forces, even gravity. The gravitational force exerted by the earth on a person is equally exerted by the person on the earth. The magnitude of the observable effect (e. g. acceleration in free fall) is different because of the extreme difference in masses.

STATIC DETERMINACY Depending on the number of forces and where they act, situations may occur where the system is statically

under- or overdetermined. Underdetermined systems allow the system to move in one or more degrees of freedom, while overdetermined systems have no unique solution for the reaction forces based on the equilibrium conditions only. The concept of static determinacy helps in detecting systems with unwanted degrees of freedom or systems under stress in overdetermined configurations.

4

DISCIPLINE-BASED EDUCATION RESEARCH (DBER)

In this chapter, research results from physics and engineering education research on conceptual understanding and active learning are presented. Furthermore, examples for the development and assessment of a specific type of RBALM are given.

In contrast to general education research, discipline-based education research (DBER) seeks to investigate the discipline-specific challenges of education, such as student difficulties in learning specific content, or issues connected to the design of curriculum and learning materials. One trend in STEM education, which began in the mid 1970s and strongly influenced education research for the following 20 years was misconception research (diSessa, 2014). Misconceptions (also known as preconceptions or alternative conceptions) can be described as

> *"[...]* conceptions *that (i) are strongly held, stable cognitive structures; (ii) differ from expert conceptions; (iii) affect in a fundamental sense how students understand natural phenomena and scientific explanations; and (iv) must be overcome, avoided, or eliminated for students to achieve expert understanding."*
>
> (Hammer, 1996)

Misconception research brought forward collections of student difficulties in the various disciplines that are still drawn upon today (e. g. Kautz (2014)[1] for introductory engineering courses, or Heron (2018) for physics courses). In physics education research (PER), for example, conceptual understanding was the main focus in the 1990s, and it still takes a large share of the research, although today, the focus of research topics has diversified as the topics go beyond an "intellectual" category, as Heron (2018) phrases it, and now also frequently address "affective and individual" affairs as well as "social and cultural" aspects. Also, the research methods have evolved. In contrast to Europe, where education research largely emerged from the education sciences also for higher education, DBER was initially dominated by researchers from within the disciplines in the US. Consequently, the research paradigm was very quantitative and focused on hypothesis testing and controlled experiments. Over time, the overall attitude towards qualitative research methods has changed towards greater acceptance, and mixed-methods designs (which are also employed in one part of this dissertation) have become more common.

1 in German

Naturally, engineering education research (EER) has also adopted research strategies and methods frequently used in PER, such as misconception research (e. g. Brown et al., 2018; Brose and Kautz, 2011), interviews (e. g. Montfort et al., 2009), pre-/post-testing (e. g. Bernhard, 2000), or longitudinal studies (e. g. Felder et al., 1998). These and other approaches slowly replace invalid methodologies that assess effectiveness of instruction merely by means of student contentment (Deslauriers et al., 2019). The shift of the US engineering accreditation agency ABET to outcome-focused criteria in the 1980s and 1990s as well as major funding by the National Science Foundation is seen as a strong catalyst for the development of engineering education as a field of research (Johri and Olds, 2013). It is hence a very young discipline that is still in a phase of self-discovery and changing quickly. Johri and Olds (2013) note that the frameworks used by EER are largely borrowed from general education research which is a sign for immaturity. But the field grows quickly as the variety of topics and methodologies presented at the leading conferences such as REES, SEFI, or ASEE shows.

4.1 CONCEPTUAL UNDERSTANDING

As there is a wide agreement among discipline-based education researchers on the importance of conceptual understanding in the STEM disciplines, it is often used as a measure for the effectiveness of instruction (McCray et al., 2003, e. g. p. 10). However, the definitions of conceptual understanding are diverse. Often, it is defined by setting it in contrast to other types of knowledge, e. g. procedural knowledge or abstract perspectives:

> *"Procedural learning is the ability to execute action sequences to solve familiar problems, procedural transfer is the ability to extend known procedures to novel contexts, and conceptual knowledge is understanding principles governing a domain and the interrelations between units of knowledge in a domain (e.g., Bisanz & Lefevre, 1992; Greeno, Riley, & Gelman, 1984; Rittle-Johnson et al., 2001)."*
>
> (Rittle-Johnson, 2006, p. 2)

> *"Conceptual understanding is often associated with intuition instead of knowledge because is it so much more internal; you don't remember something you understand conceptually, it is just true. [...] Conceptual understanding (people's personal explanations of how and why the world works) is knowledge in context, and is therefore more transferable than computational ability."*
>
> (Montfort et al., 2009, pp. 111-112)

Other definitions focus on the organizational structure of conceptual understanding. For example, Streveler et al. (2014) define it as a collection of an individual's

> *"[...] concepts, beliefs, and mental models, where the following definitions apply:"*
>
> - "Concepts *are pieces or clusters of knowledge, for example, 'force,' 'mass,' 'causation,' and 'acceleration.'"*
> - "Beliefs *are propositional relationships between concepts, for example, 'a force on a mass causes acceleration.'"*
> - "Mental models *are groups of meaningfully related beliefs and concepts that allow people to explain phenomena and make predictions; for example, an expert dynamics instructor would use her mental model of Newtonian physics to predict an object's motion"*
>
> (Streveler et al., 2014, p. 83)

Still others choose an outcome-based description of the nature of conceptual understanding. For instance, Edström paraphrases student conceptual understanding as the ability to

> *"[...] explain matters in their own words, interpret results, integrate knowledge from different courses and apply it to new problems."*
>
> (Edström, 2012, p. 4)

According to Edström, the frequently observed lack of this ability is not surprising, as students do not practice it enough in a typical engineering curriculum.

While each of the different attempts to define conceptual understanding that were quoted above is certainly useful for different research questions, neither the organizational structure of conceptual understanding (Streveler et al., 2014) nor the macroscopic outcome-based perspective on the effectiveness of entire study programs that would call for the ability to integrate knowledge from more than one course (Edström, 2012) is of interest in this dissertation. Instead, this dissertation makes use of the fact that having a functional conceptual understanding allows to make general statements about a system without requiring specific parameters to perform calculations. As such, conceptual understanding becomes an observable construct, which will be exploited in the measurement of the same. Here, it is hence rather the contrast to procedural knowledge (Rittle-Johnson, 2006) that is most useful in characterizing conceptual understanding.

Unfortunately, lack of conceptual understanding by students often remains undetected as assessment often focuses on procedural

knowledge, as exemplified for instance by Mazur (1997). Mazur correlated student scores on two tasks addressing DC circuits, one conceptual, requiring qualitative reasoning, and the other one conventionally quantitative. The conceptual problem features a simple DC circuit with three bulbs, one current source and a switch. The students are asked how several quantities in the circuit change qualitatively, when the switch is closed. The supposedly (as rated by instructors) more difficult conventional problem asked to determine a specific current and voltage quantitatively in a more complex circuit, which requires setting up and solving a system of equations using Kirchhoff's laws. Both problems were given on the same exam. The results showed that students are often able to answer the supposedly more difficult quantitative problem, while they could not answer the supposedly easy conceptual question. The reverse case was not observed very frequently. Although this is an example from the physics curriculum, it applies to engineering as well. It shows two things: (1) a more complex task is not generally more difficult, and (2) one cannot conclude from good student performance on conventional quantitative questions that they have achieved to grasp the content on the conceptual level. If both, "[...] conceptual understanding and individual problem-solving skills are necessary outcomes of engineering education [...]" (Edström, 2012, p. 5), students must also be assessed on both and given the chance to practice both. Similar arguments are brought forward by McDermott (2001) who concludes from years of PER experience that

> "*1. Facility in solving standard quantitative problems is not an adequate criterion for functional understanding. [...] 2. Connections among concepts, formal representations, and the real world are often lacking after traditional instruction. [...] 3. Certain conceptual difficulties are not overcome by traditional instruction.*"
>
> McDermott (2001)

All of these aspects illustrate the importance of conceptual understanding for engineering students and make conceptual understanding a suitable indicator for assessing the effectiveness of instruction.

4.2 ACTIVE LEARNING

The majority of studies on various active learning methods show a positive effect of an active learning strategy format on gaining understanding in contrast to traditional instruction. Smith et al. (2005) provide an excellent overview on classroom-based active pedagogies and their effect on student learning. While there is (once again) not a unique definition of active learning, the definitions found in literature are largely similar. For example, Freeman et al. (2014) state the following definition:

> *"Active learning engages students in the process of learning through activities and/or discussion in class, as opposed to passively listening to an expert. It emphasizes higher-order thinking and often involves group work."*

Using the term "interactive engagement", Hake (1998) defined active learning as methods

> *"designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors"*

Both definitions address the desired outcome (higher-order thinking/conceptual understanding), the activity (involving hands-on activities and discussion), and the social element (group work/peers).

An early and well-known large-scale study on the effectiveness of interactive teaching was conducted by Hake (1998) in the physics context. He analyzed data based on responses from over 6000 students to concept inventories used as pre- and post-test in various physics courses. Instruction in these courses was categorized as either "interactive engagement (IE)" or "traditional (T)". Hake defines traditional courses as "relying primarily on passive-student lectures, recipe labs, and algorithmic-problem exams". His definition of IE is given above. He used the average normalized gain g as a quantification of how much students in each course have improved their conception of force. g relates the course average absolute gain in score from pre- to post-test to the maximum possible gain, allowing for comparison of courses with different average pre-test scores. Hake showed that on average interactive instruction leads to a much higher gain than traditional instruction ($\bar{g}_{IE} = 0.48 \pm 0.14$, $\bar{g}_{T} = 0.23 \pm 0.04$).

→ Section 15.1 for details on g.

Freeman et al. (2014) showed that the conclusion that conceptual understanding is fostered more by interactive than by traditional methods is valid beyond the physics context. They conducted a meta-analysis on the results of 225 studies comparing active learning and traditional lecturing in various STEM disciplines. As a criterion for admitting studies to the meta-analysis, they defined active learning as above. For instruction to be categorized as traditional, "[...] student activity was assumed to be limited to taking notes and/or asking occasional and unprompted questions of the instructor." Performance indicators were assessments such as exams or concept inventories on the one hand, as well as failure rates on the other hand. The latter were found to be 1.5 times higher in traditional settings. The performance on the assessments was found to differ by an average effect size of 0.47 in favor of the active instruction. (When only engineering disciplines were considered, the value was comparable to this result.) The authors note that all instructors of active learning voluntarily used the method and that the results cannot be generalized to a setting where active learning instruction would be mandatory.

Von Korff et al. (2016) performed a secondary analysis of FCI and Force and Motion Conceptual Evaluation (FMCE) pre-/post-test data that had been published since the first employment of these instruments. Their results, which are based on data from about 45000 students / 600 courses, confirm previous findings on the superiority of student interactive engagement over traditional lecturing in producing high gains on concept inventories. Furthermore, they were able to show that interactive engagement is effective in various settings, independent of students' prior knowledge, class size, or type of institution.

A limitation of all these meta studies is the possibility of a selective publication bias[2]. Andrews et al. (2011), hypothesize that the positive results from active learning studies may be biased by instructor expertise, because education researchers often investigate their own courses. Data from randomly selected courses with active and traditional instruction showed no significant effect on student learning that could be attributed to the use of active learning methods alone. Instead, the consideration of misconceptions was identified to be one of the most significant factors for successful concept learning. Thus, activity may be essential for successful conceptual change but it is not sufficient. This statement is also supported by Nie et al. (2019) who present evidence that drilling students in problem-solving is not successful at changing their misconceptions towards expert-like conceptions. Their studies are situated in the Chinese secondary school context but the identified misconceptions were identical to the ones known from the US higher education context, indicating that misconception research results are in principle transferrable to other national contexts and educational levels.

Another disadvantage of meta studies is that the categorization into only two forms of instruction, i. e. (inter)active or traditional, is very coarse for the multitude of different interactive learning scenarios. This shows for instance in Hake's data in the larger variance of the gains in the interactive category and in an overlap with the traditional category, i. e. not every interactive course achieved higher gains than the traditional courses. While instruction design should be informed and guided by such results, each implementation of instruction must still be assessed on its own for its effectiveness.

All of these results and many more support the statement made by Finkelstein and Pollock (2005) that "[r]esearch-based materials in both subject content and pedagogical approach are essential". RBALM can take many different shapes. McDermott and Redish (1999) provide a list of research-based instructional materials for physics instruction developed in the 1990s.

2 In case of Hake (1998), the data was not necessarily published in a journal but instructors themselves decided whether or not to provide their data.

One specific set of instructional materials are the *Tutorials in Introductory Physics* developed by the Physics Education Group at the University of Washington throughout the 1990s and first published a few years later (McDermott and Shaffer, 1998). These materials and their implementation act as a role model for the materials investigated for effectiveness in this dissertation. The underlying concepts, the development process and prior research on their effectiveness is discussed in the following sections.

4.3 TUTORIALS

Arnold Arons, a pioneer in the promotion of PER in the US was an advocate for Socratic dialogue, which he defines in the classical sense as "a conversation in which an interlocutor asks his partner a sequence of questions eliciting a line of reasoning or inquiry without making didactic interpolations" (Arons, 1981). After coming to the University of Washington in 1968, where he worked with pre-service teachers, he began research on intellectual development, concept formation, and reasoning abilities of students at college level (Minstrell, 1981). Arons found that students' mental models often differ from the one held by the instructor and that students often struggled with various concepts, for example, ratios and scaling, discriminating time instants from time intervals, and also with the concept of force and free-body diagrams (Arons, 1981). The results of his research contributed largely to his books on how to teach physics (e. g. Arons, 1990) which are seen as a valuable resource for Physics teachers (Blanton, 2001). Inspired by Arons' findings, Lillian McDermott established the *Physics Education Group (PEG)* in the 1970s which adopted and refined many of Arons' research and instructional approaches. Together with Peter Shaffer, McDermott created the *Tutorials in Introductory Physics* (McDermott and Shaffer, 1998). These make extensive use of Socratic dialogue as they exist in the form of collaborative-group worksheets intended to be used in multiple small groups of three to four students under the supervision of teaching assistants (TAs) who have been trained to engage the students in Socratic dialogue. They are often referred to as "qualitative" worksheets because they do not involve extensive calculations but focus on understanding the concepts.

An instructional strategy that can be found in many Tutorial worksheets is "*elicit, confront, resolve*" (McDermott, 1991; Finkelstein and Pollock, 2005; Heron, 2018). In the context of a specific physical situation, the Tutorials often ask students to make a prediction, for example, how a given quantity will change, thereby *eliciting* the students' current conceptions. These prompts intentionally address common student difficulties (often identified by prior research) so that in many cases, a specific incorrect prediction is made by the students. The design of the worksheet then guides the students towards a contra-

diction in their own reasoning or *confronts* them with contradicting experimental results. Guided by further questions, the students are then expected to *resolve* this cognitive conflict. Other common didactic elements of the Tutorials exhibiting the same strategy are, for example, fictitious student statements that involve known misconceptions, which are then to be discussed by the students.

Since the publication of the *Tutorials in Introductory Physics*, similar instructional materials, mostly for more advanced topics in physics, have been developed by other researchers. Among others, these include the *Activity-Based Tutorials* (Wittmann et al., 2004, 2005) and (Steinberg et al., 1997), the *Intermediate Mechanics Tutorials* by Ambrose and Wittmann (2007), *PER-Based Tutorials for Quantum Physics* by Ambrose (2014), and the *University of Pittsburgh E&M Tutorial Series* by Singh and University of Pittsburgh Physics Education Research Group (2012). At Hamburg University of Technology (TUHH), the Engineering Education Research Group has used a similar approach to develop *Tutorials* for other subjects in the introductory engineering curriculum. Apart from the *Tutorials in Engineering Mechanics* (Kautz et al., 2018) investigated in this dissertation, materials were developed for Electrical Engineering (Kautz, 2010) and Programming (Timmermann and Kautz, 2017) (all in German).

4.3.1 *Constructivism and conceptual change as theoretical framework for Tutorials*

Cognitive scientists have developed multiple learning theories during the last century. They are often categorized into three main theories in literature: behaviorism, cognitivism and constructivism[3]. While constructivism is the dominant learning theory these days and is used as the basis for many instructional or research designs, the other theories can still provide a valuable framework for some mechanisms of learning, for example, sensorimotor skills. Conceptual understanding, however, may be best described by constructivist learning theories.

The central assumption in constructivism is that learners construct knowledge for themselves. As in cognitivism and in contrast to behaviorism, the internal structure of knowledge and the processes of knowledge building are considered, but there are different constructivist views on the precise origin and nature of knowledge that are more or less extreme (Schunk, 2012).

1. The *exogenous* perspective assumes that there exists an objective truth and by interacting with the outside world individuals reconstruct this truth. Their knowledge is seen as a mirror of the external world, which varies in accuracy among novices and experts.

3 Strictly speaking, constructivism is not a learning theory but an epistemology (Schunk, 2012, p. 230).

2. The *endogenous* perspective sees knowledge as being created within the individual from old knowledge, not mirroring external structures.

3. The *dialectical* perspective lies in between these two extremes. It emphasizes the social aspect of knowledge building and the influence of the environment. Mental conflicts are primarily triggered by interaction with other persons or the environment.

Piaget's theory of cognitive development (Piaget, 1970) is presented by Schunk (2012) as an example for endogenous constructivism. However, in the author's opinion, most aspects of Piaget's theory show a closer connection to dialectical constructivism in the sense that it emphasizes interaction with the environment. According to Piaget's theory, the desire for *equilibration* is the core driving force for cognitive development, apart from biological maturation and experience with the physical and social environments. Interaction with the environments leads to the reception of new information which can create a conflict or disequilibrium with the individual's existing knowledge. The human mind strives to reestablish equilibrium, which can be achieved by two means: Borrowing concepts from biology, Piaget theorizes that the new information is either *assimilated* to existing knowledge structures, thereby strengthening them, or the existing knowledge structures are *accommodated* to fit the new information.

It may be the internal nature of the desire for equilibration that gives the theory an endogenous character, but any disequilibrium arises from interactions with the environment, which speaks for a dialectical character.

> *"Knowledge, then, at its origin, neither arises from objects nor from the subject, but from interactions - at first inextricable - between the subject and those objects."*
>
> (Piaget, 1970, p. 704)

This statement implies that students must be active ("inter*actions*") because (objective) knowledge is *constructed* from the interactions with objects. The initially limited and primitive actions evolve into more complex physical or mental operations when moving through the *stages* of cognitive development: sensorimotor (birth to 2 years), preoperational (2 to 7 years), concrete operational (7 to 11 years), and formal operational (11 to adult). According to Piaget, these stages are discrete and passed in a fixed sequence. The approximate ages at which children enter a stage as well as the independence of the theory from specific disciplines has often been criticized (diSessa, 2014). These "flaws", however, are what makes the remaining aspects of the model still relevant for higher education as even adults may not be on the same stage across all topics. For example, students may understand the concept of reaction forces in Statics at the concrete operational level (e. g. "When the loads are quantitatively given, I sum

up all the forces and moments. The sums must be zero, and I solve the equations for the unknown reaction forces."), or at the formal operational level (e. g. "I can explain which types of support can exert which type of reaction and calculate the reaction forces, which are always imposed by the equilibrium conditions as a reaction to the applied loads.").

Schunk (2012) lists four implications for instruction that result from Piaget's theory of cognitive development. The *Tutorials* and their implementation address all of these points:

1. "*Understand cognitive development.*" Instructors should understand the stages of cognitive development and adapt instruction according to the individual student's stage. The *Tutorials* were developed based on research of typical student thinking and misconceptions. They are thus designed to fit the average stage to be expected in the population. Furthermore, the Socratic dialogue between students and the instructor (or TA) in the setting of the *Tutorials* provides the instructor with an opportunity to elicit the individual stages and react accordingly.

2. "*Keep students active.*" The "minds-on" and sometimes "hands-on" design of the *Tutorials* strongly promotes active student behavior.

3. "*Create incongruity.*" "Learning occurs, then, when children experience cognitive conflict and engage in assimilation or accommodation to construct or alter internal structures" (Schunk, 2012, p. 238). The *elicit, confront, resolve* theme is designed to create a cognitive conflict and have the students resolve it themselves.

4. "*Provide social interaction*" because it fosters cognitive development. The peer group discussions are a key part of the *Tutorials*. Learning from peers and discussing potentially conflicting viewpoints provide more opportunities for resolving disequilibrium between their own mental structures and environmental input.

Judged from within the framework of constructivist theory, the Tutorials thus have the potential to be very effective learning materials. The theoretical framework of the Tutorials is furthermore strengthened by considering conceptual change theory. There are several theories of conceptual change which all agree that preconceptions affect the learning process (Streveler et al., 2014). The *Tutorials in Introductory Physics* were developed without a strong commitment to either one of the streams of conceptual change theory (Heron, 2018), but conceptual change theory reflects the characteristic *elicit, confront, resolve* theme and supports a Socratic dialogue strategy. For example,

based on Piaget's idea of equilibration, Posner et al. (1982) postulate four conditions to be fulfilled before an individual will replace a (mis)conception by a new conception (*accommodation*): The initial condition that serves as an impulse to challenging existing conceptions is *dissatisfaction*. As long as conceptions are suitable to explain a phenomenon, there is no benefit for a human being to make an effort to challenge them. Furthermore, the new conception must be *intelligible*, (i. e. the essence of it must be understandable), *plausible*, (i. e. the reasoning must make sense), and it should be *fruitful* (i. e. it may explain more phenomena than just the one at hand).

These presumed requirements for successful learning may explain common observations of the limited success of instruction in STEM disciplines, as for example the following from physics:

> *"[T]he traditional forms of instruction seem to be inadequate for helping most students develop a functional understanding of basic topics in physics. Hearing lectures, reading textbooks, solving quantitative problems, seeing demonstrations, and doing experiments often have surprisingly little effect on student learning."*
>
> McDermott (2001)

Even if the ideas presented by traditional instruction are indeed perceived as intelligible and plausible by students, but possible conflicts with their current conceptions remain implicit, the condition of dissatisfaction is not fulfilled and conceptual change is unlikely to happen. Effective instruction must therefore begin by making the learner aware of their existing ideas about the subject matter (elicit) and help them recognize which of these are consistent with what is being taught (confront and resolve). The research-informed Tutorial worksheets in combination with activating group discussions and Socratic dialogue aim at fulfilling these conditions by creating situations and an environment that foster conceptual change.

4.3.2 *Tutorial development process*

Discovering common misconceptions among students that can help create the state of "dissatisfaction" is the initial step and constitutes a major part of the development process of the Tutorials. The first indication for a potential difficulty might arise informally, e. g. from teacher experience in the classroom, by carefully listening in on student discussions, or when grading course examinations. At this stage, a hypothesis for the existence of a misconception may be developed, which can then be tested in student interviews or through written questions that generally focus on the behavior or properties of a specific physical sample system. Interviews are limited to small sample sizes, but they allow to probe student thinking more deeply than written tests. In particular, a semi-structured interview approach provides

the opportunity for the discovery of further misconceptions. Recurring patterns of thinking that appear in the interviews are likely to be also widely shared by the entire population. Written tests designed on the basis of interview results can later help quantify the frequency of the identified misconception.

Often, interview questions that effectively elicit student thinking and trigger specific misconceptions also serve well as an instructional tool and can therefore be used as a starting point in the construction of a Tutorial worksheet. A step-by-step analysis of the thinking processes required to arrive at an understanding of a given problem situation can be used to outline the structure of the planned Tutorial. The detailed design then involves the use of various didactic elements/building blocks such as predictions, observations, or generalizations.

The development of the Tutorials is an iterative process. Feedback from instructors and teaching assistants is collected on a regular basis and used to improve individual worksheets. In addition, results from open-ended questions on course examinations may point to student difficulties that persist even after instruction with Tutorials, leading to further modification.

In order to illustrate this iterative process the following example from the instructional context investigated in this dissertation is presented: Administrations of a conceptual post-test on statics suggested that the concept of static equivalence is a major conceptual difficulty for students. This finding sparked an in-depth investigation using individual student interviews (Brose and Kautz, 2011). Strong evidence was found for the misconception that single forces and moments can, in principle, be (statically) equivalent; specifically, that a moment M at a given point P can be replaced by a single force $\vec{F}$ acting at distance d from P as long as the force causes the same moment about point P (i.e., $\vec{M} = \vec{d} \times \vec{F}$).

The Concept Assessment Tool for Statics (CATS), see Chapter 6.

The insights gained in these and similar interviews were used to create Tutorial worksheets that address the identified difficulties. Figure 3 shows an example how the misconception is addressed by the Tutorial. In part a, the students' own thinking is *elicited* by asking them to agree or disagree with a statement by a fictitious student who apparently holds this misconception. Having to justify their answer helps making their conception explicit. In parts b and c, students having the same incorrect idea will be *confronted* with a contradiction between the misconception and the concept of static equivalence as neither the resulting forces nor the resulting moments about Q (or any point other than P) are equivalent in the two situations. Instead of simply presenting a correct answer, part d gives the students the opportunity to *resolve* the conflict by themselves (while teaching assistants are available for help, if required).

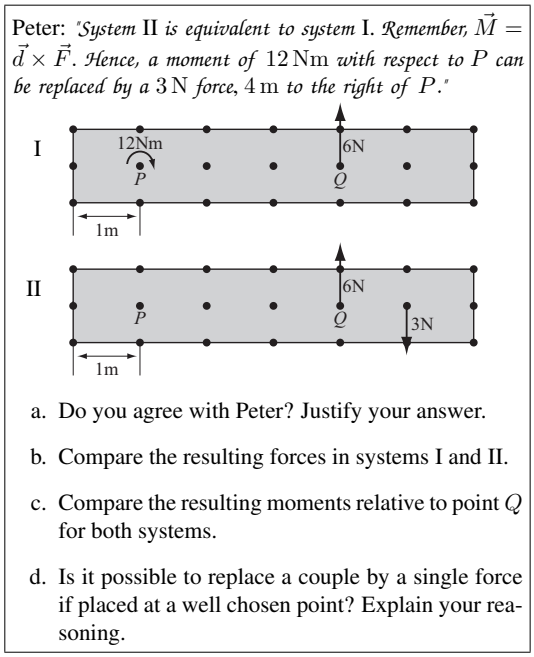


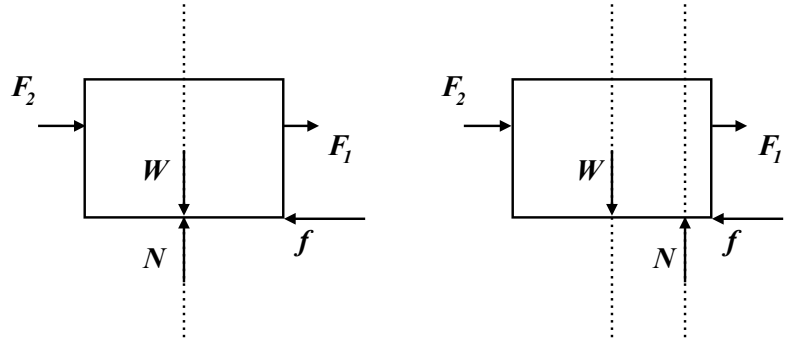
Figure 3: Excerpt from the Tutorial worksheet on equivalence of force systems, designed to address misconceptions about the interchangeability of forces and couples.

4.3.3 *Differences between Tutorials for physics and engineering instruction*

The *Tutorials for Engineering Mechanics* were developed in part based on the same misconceptions that were found among physics students, while considering the characteristics of engineering. Hence, the question may arise why engineering specific Tutorials must be developed. The following paragraphs illustrate how the differences between physics and engineering instruction, which have been addressed in Chapter 1, are reflected in the development of the Tutorials by comparing two worksheets on the same topic, one designed for physics and one designed for engineering. The example for a Physics Tutorial addressing rigid-body mechanics is the worksheet *Equilibrium of Rigid Bodies*, which considers student difficulties with the concepts of center of mass, forces, moments and equilibrium. Ortiz et al. (2005) describe the misconceptions identified in student interviews, using simple physical systems of an object balanced on a single-point frictionless support (such as a baseball bat balanced horizontally on a finger). For example, the center of mass is considered to literally divide the object into two halves of equal masses (which is not the case for objects of non-symmetric shape or heterogeneous density). This misconception is closely linked to the belief that the *forces* on either side of the support must balance instead of the *moments*. Ortiz et al. (2005) also found that students have difficulty with the simplification

of considering a gravitational force as acting at the objects' center of mass, especially if part of the object extends to the other side of the support. The Tutorial that was developed based on these misconceptions involves hands-on experiments with a T-shaped object on a pivot. It starts by helping students understand why objects supported at their centers of mass remain at rest. Subsequently, it helps students realize that balance depends not only on the mass on either side of the pivot but also on its distribution. This idea of balanced moments is then applied and practiced, by reasoning that the greater mass must have its center closer to the pivot. Finally, the aspect of equilibrium of objects supported at a tilted angle is addressed.

The *Forces and Moments* worksheet by Kautz et al. (2018) addresses the same general topic. First, a system consisting of two or more masses resting on a rigid, massless beam supported by a fulcrum at its center is inspected. It starts similarly to the previously described Tutorial by helping students realize that the balanced state depends on both, the masses and their positions on either side of the support, by asking students to predict the effect of changing either the mass or its position on only one side of the support. The students then continue to inspect the directions and magnitudes of the moments on either side of the support for different combinations of masses and positions. From this, they induce the equilibrium condition $\sum \vec{M} = 0$, and then go on to inspect the effect of choosing a different point of reference. The following addresses the special characteristics of engineering: applying force and moment equilibrium conditions *at the same time*. The physical situation describes two people trying to slide a crate over a high friction ground, which prevents the crate from slipping. The people exert horizontal forces in the same direction at shoulder's height. The students draw a FBD of the crate and examine the moments caused by every single force in their FBD about the lower corner of the crate. Most students assume correctly that the vertical forces, i. e. the weight force and the normal force by the ground, are equal in magnitude, but most incorrectly assume that they are aligned so that their combined effect on the sum of moments is zero (see Figure 4 (a)). The Tutorial leads these students into a cognitive conflict once they realize that the moment caused by the horizontal forces exerted by the people and the friction force remains unbalanced. This conflict is resolved by helping them realize that the normal force (distribution) shifts due to the moment exerted by the horizontal forces (see Figure 4 (b)). The conditions for tipping are examined in a thought experiment, assuming that the people increase their effort to move the crate, which still does not slip. Finally, moment equilibrium of the crate is again examined with a qualitatively different point of reference, leading to the conclusion that moment equilibrium holds for all reference points, even those external to the object.



(a) incorrect FBD (b) correct FBD

Figure 4: Free-body diagrams describing a situation from the Tutorial *Forces and Moments*. Students are guided to see that the normal force N shifts as a reaction to the moment introduced by the horizontal forces.

The wholistic consideration of *all* equilibrium conditions at once is one of the main aspects in which the *Forces and Moments* worksheet (Kautz et al., 2018) designed for engineering students differs from the *Equilibrium of Rigid Bodies* worksheet (Ortiz et al., 2005) designed for physics students. The fact that a change in horizontal forces can affect the vertical forces through the resulting moments and the related concept of equilibrium stability aim at a core task of engineering: designing stable systems. Thus, such differences between the disciplines as the one illustrated by this example justify the development of discipline-specific Tutorials even if the misconceptions to be addressed are the same in existing Tutorials.

4.3.4 *Prior research on the effectiveness of Tutorials*

Assessment of the effectiveness of Tutorials has been carried out in different ways. In the process of developing the *Tutorials in Introductory Physics* (McDermott and Shaffer, 1998, and subsequent editions) a vast amount of research was produced on student misconceptions (in the manner described above) and on the effectiveness of individual Tutorials by members of the Physics Education Research Group at University of Washington (e. g. Wosilait et al., 1998; Ortiz et al., 2005; Shaffer and McDermott, 2005; Cochran and Heron, 2006; Loverude et al., 2010; Close et al., 2013). In these studies, the effectiveness is usually measured by short open-ended conceptual questions implemented as pre- and post-tests to the respective Tutorial worksheet. Post-Tutorial results often show 70 % or more correct answers while at most 50 % (and often considerably less) tend to be reached with traditional instruction only. For assessing the success of a worksheet, the performance of graduate TAs in the subject on these same questions is generally taken as a reference (e. g. Ortiz et al., 2005, p. 552).

Finkelstein and Pollock (2005) were not involved in the development of the Tutorials, but were able to replicate the results with similar open-ended questions. They showed that students obtained a high level of conceptual understanding, along with a high average normalized gain (see Equation (15)) on relevant concept inventories. In their study, they also consider the influence of the environmental structure at the course and department levels on the learning process, concluding that the implementation of educational reforms developed elsewhere must go along with careful adaptations to fit the local boundary conditions. Otherwise, the reform is likely to fail, a case which was observed by Riegler et al. (2016) in the German higher education context. They identified several conceptual and structural mismatches which hindered a successful implementation of the Tutorials. Amongst other reasons, the scaffolding by the Tutorials was perceived as too strong by the students. Furthermore, many of the misconceptions targeted by the Tutorials were not common among the students, most likely because they already had a high level of conceptual understanding from high school physics. Finally, the authors noticed deficiencies in TA preparation and buy-in.

The importance of these latter factors is emphasized by Koenig and Endorf (2003) and Koenig et al. (2007). Their results suggest that the success of the Tutorials depends strongly on the TAs and their ability to engage the students in Socratic dialogue. To determine the crucial elements of a successful Tutorial implementation, Koenig et al. (2007) used open-ended questions to investigate the effectiveness of four different implementations of the Tutorials, ranging from a traditional lecture, through students working alone and working in small groups, to the recommended group setting, properly assisted by a TA. The recommended setting led to substantially better student understanding than any of the other three settings.

The results presented by Kryjevskaia et al. (2014) seem to contradict this finding by suggesting that the setting specified by the developers might actually be more flexible for adaptation than assumed. Tutorials used in an active-lecture setting resulted in similar gains on open-ended conceptual pre-/post-test questions as in the recommended setting. However, in both of the active-lecture settings described there, a number of favorable conditions which address the central aspects of Tutorial implementation were satisfied. For instance, students were given the opportunity to think about critical questions before instruction on a given topic (resulting in an "eliciting" phase as described above). Furthermore, the instructors were very familiar with typical student reasoning about the topic (through their own research). Therefore, they could respond adequately to student statements indicating specific difficulties (misconceptions), which likely led to meaningful class discussions. In any case, the challenges for Tutorial TAs (or

instructors) remain demanding. They are described by Heron (2018) as

> *"exercis[ing] judgment and creativity in responding as groups tackle thorny issues: When to ask an additional question to sharpen the issue? When to ensure that all viewpoints have been properly heard? When to step back and allow conversation to proceed organically?"*

While the previously mentioned studies all involved Tutorial implementations at the college level, Benegas and Flores (2014) studied the effectiveness of two Tutorials on DC circuits in Argentinean high schools, considering also long-term effects. The students' conceptual understanding (as measured by a widely-used concept inventory) was systematically higher for students in the classes using Tutorials than in the control group, not only immediately after instruction but also one year later. The control group's test results dropped back to pre-instruction level, indicating that the traditional instruction was ineffective in improving conceptual understanding on a longer time scale, while Tutorials not only led to better understanding but also to more sustainable conceptual change.

5 DIAGNOSTIC TESTS AND CONCEPT INVENTORIES

> *"[A concept inventory (CI) is] a multiple-choice instrument designed to evaluate whether a person has an accurate and working knowledge of a concept or concepts."*
>
> (Lindell et al., 2006)

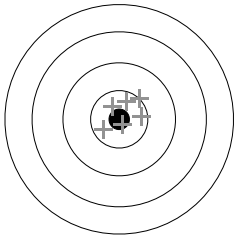
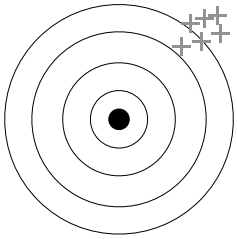
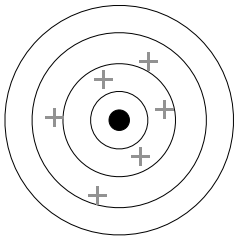
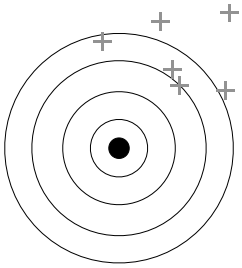
CIs are a form of assessment frequently used by discipline-based education researchers and instructors for various purposes, mostly to give summative or formative feedback to the students, to learn about students' misconceptions or to evaluate and compare instruction (Engelhardt, 2009). Unlike midterm or final exams, they are "highly focused on a small set of key concepts within a limited academic content domain" (Jorion et al., 2015). In this respect, Engelhardt (2009) warns against using them as a replacement for final exams, but acknowledges that "these tests can be used to determine how well a new teaching method or curriculum helps to remedy these known misconceptions and improve the quality of teaching".

In this chapter, the concepts of validity and reliability as well as a general strategy for the development of CIs is introduced and examples of CIs on concepts that are relevant to engineering mechanics are presented.

5.1 VALIDITY AND RELIABILITY

The concepts of *validity* and *reliability* are essential to the quality of any measurement instrument. They are often visualized by an analogy to archery (see Table 1). Validity addresses the issue of measuring the *intended* construct, while reliability refers to the degree of accuracy of the measurement. The "bull's eye" on the target symbolizes the construct to be measured. Each mark represents one measurement. Reliability is characterized by the variance in repeated measurements, independent of which construct is actually measured. Low variance stands for high reliability and vice versa. Validity is characterized by the degree of how much the measurements are centered around the construct to be measured, so that (under consideration of a random measurement error) the actually measured construct matches the desired one. Unlike Table 1 suggests, validity is not independent of reliability. A certain degree of reliability is a prerequisite for validity (Moosbrugger, 2012, p. 120) as no construct can be measured with an instrument that produces pure measurement error.

Table 1: Visualization of the validity and reliability concepts.

Reliability	Validity: high	Validity: low
high		
low		

There are several aspects which might pose a threat to validity, e. g. if the items are misunderstood, or the construct tested by an item is not related to the overall construct targeted by the test. The statements "the test is valid" or the term "validity of a test" can be found in many publications. Care must be taken as they falsely suggest that validity is a property of a test. Instead, validity refers to the interpretation of test scores for a predefined purpose in a certain context. "The test is valid" is a mere shortcut for saying that the interpretation of the test scores for a predefined purpose is valid, if the test was administered in a predefined way. Some authors may use it simply as an abbreviation, others may actually hold the misconception which is reflected in these phrases: that validity is a property of a test, independent of purpose, test population, time, or other conditions. For the sake of brevity, the shortcut will also be used in this dissertation. Terms such as "content validity", "construct validity", "criterion validity" or "face validity" are frequently used to describe the different aspects which the collected evidence may address:

FACE VALIDITY can be quickly evaluated "at a glance" as it addresses the aspect whether a test superficially appears to measure the intended construct. For example, items which show representations of electric circuits and ask for the magnitude of a current are unlikely to measure understanding of engineering mechanics at face value. In contrast, items which, for instance, show mechanical systems at rest and ask for equilibrium conditions are more likely to be face valid.

CONTENT VALIDITY goes deeper. It relates to the aspect that necessary content related to the measured construct is addressed. For example, understanding of statics may not be sufficiently assessable if only moments were considered as loads and forces were omitted. In contrast, items testing the understanding of the fundamental statics concepts in three-dimensional systems in addition to two-dimensional systems may not be required.

CRITERION VALIDITY can be established by comparing the results to a gold standard measurement, if available. Often, this is not the case, but a positive correlation with results from instruments measuring similar constructs are taken as evidence. Final exams, for instance, measure conceptual understanding but also procedural problem-solving skills. Thus they are similar instruments, but perfect correlation cannot be expected.

CONSTRUCT VALIDITY touches on the aspect whether the test actually measures what it claims to measure, or whether individual items or the entire test capture other constructs.

Care must be taken that these different aspects of validity are not mistaken as different possibilities of how a test can be valid. They must all be fulfilled for a test to "be valid", as the following definition suggests:

> *"Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use."*
>
> (American Educational Research Association et al., 2014, p. 14)

According to this definition, the purpose of or proposed use for a test is an essential aspect of validity. It should be considered already in the development phase, as developing an instrument for multiple purposes inevitably leads to compromises (National Research Council, 2001). For instance, one test instrument cannot serve equally well as a diagnostic tool for the specific difficulties among weak performers on the one hand, and as placement test for an elite education programme on the other hand (Crocker and Algina, 1986, p. 67), or other purposes holding possibly serious consequences for the individual. Similarly, using class average test scores on a CI in order to compare the effect of different instruction on conceptual understanding is different from using it as a formative assessment instrument for the individual. In case of the former, instructors or education researchers may also be interested in the frequency of certain misconceptions among the students in order to design appropriate instruction, which again sets a different focus on the investigation of validity. Likewise, which level of reliability is required depends largely on the purpose

in two aspects: making statements about individuals vs. making statements about groups and high stakes vs. low stakes. When making high stakes decisions about individuals based on a test result, such as admission tests to university, reliability standards should be higher than when making low stakes decisions. For group ratings, errors tend to level out, making the average values more reliable than individual scores. For the purpose of comparing cohorts with different instruction, a moderate reliability may hence be acceptable.

Consequently, the interpretation of scores and the purpose of a test must both be made explicit before any investigation on validity can be commenced (American Educational Research Association et al., 2014).

5.2 DEVELOPMENT OF CONCEPT INVENTORIES

The design process of CIs is usually research-based, with the goal of creating an instrument that allows reliable measurements and valid interpretations of the test results. CIs are most often designed as multiple-choice instruments because of the many advantages of this format, most importantly highly objective scoring and the possibility of automated and thus fast evaluation, even when administered to large populations. These assets are important for standardized instruments as they are meant to be widely implemented, which becomes more likely if the administration is cheap. Furthermore, multiple-choice instruments allow for easy application of test theory methods for validity and reliability checks. With properly designed multiple-choice items, even higher-level thinking can be measured[1].

This efficiency comes at the price of giving up the opportunity to react to students' responses individually in order to probe more deeply and to discover yet unknown misconceptions. The multiple-choice question (MCQ) format of CIs is furthermore prone to guessing behavior, and the collected data in form of a selected response is not rich enough to ensure that students have sufficient understanding to explain why the selected response is correct and the others are incorrect (Boles et al., 2015). Developers try to contain this source for false positive responses on individual items by validating student reasoning for example through interviews or analyzing open-ended versions of the items. In addition, the composition of the tests is often such that multiple items measure the same concept which increases reliability of the total test score. Some developers prefer two-tier multiple-choice instruments, where the first tier asks students to commit to a response, the second tier asks for a justification, also in multiple-choice format (Treagust, 2012). Thereby, student reasoning is made explicit, whereas multiple-choice diagnostic tests without the second tier must rely on

1 Haladyna et al. (2002) present a taxonomy of multiple-choice item-writing guidelines.

the implicit mapping of distractors to misconceptions. On the downside, high-quality two-tier items are often very difficult to design as, ideally, all statements given as second-tier options should be correct in themselves, with only one being the actual justification for the response to the item (Timmermann and Kautz, 2015). Recent research in textual analysis may soon irradicate this problem: Goncher and Boles (2019) used CIs enhanced by short explanations from the students, which were analyzed for guessing by means of textual analysis.

Another problem, for which the root cause is still unknown is a frequently observed gender gap in favor of males on various concept inventories which is much lower on other measures such as exams. For example, the FCI and the FMCE both have an average gender gap of 12 to 13 % compared to 0 to 4.5 % on regular physics exams (Madsen et al., 2013). While the reasons for this gap are unclear the implications for research are evident: Care must be taken if effectiveness of instruction is investigated by comparing groups of different gender compositions.

Crocker and Algina (1986) lay out a detailed ten-step process of standardized test design as follows (slightly reworded here to apply to concept inventories):

1. Define the purpose of the test.
2. Identify the concepts to be tested.
3. Define the test specifications, e. g. format and number of items per concept.
4. Create an initial pool of items.
5. Have items reviewed by experts (and revise as necessary).
6. Hold preliminary item tryouts (and revise if necessary).
7. Field-test the items on a large representative student sample.
8. Conduct item analysis (and remove items that do not meet previously established criteria)
9. Design and conduct reliability and validity studies for the final test.
10. Develop guidelines for administration, scoring, and interpretation of the test scores.

These recommendations were written in 1986 but they are still applied without substantial adaptations in more recent literature (e. g. Engelhardt, 2009; Adams and Wieman, 2011). It becomes evident that the development of a CI is a lengthy process. Steps 5 and 6 are often done in interview formats which are very time-consuming and step 7, field-testing, is often only sensible during a certain time of the

semester or even year. On top of that, time for revision loops must be accounted for. Hence, whenever possible, existing instruments should be adopted instead of creating one's own tests in a quick shot. In that case, it must only be ensured that the instrument is suitable and validly interpretable for the given context, i. e. step 9 may have to be repeated.

Test theory provides a framework for steps 8 and 9. Classical Test Theory (CTT) is based on the theory of measurement errors. The central assumption in CTT is that the observed test score X of a person P on an item i is composed of the true score τ and a measurement error ε. Reliability is defined consistently with this assumption as the fraction of true score variance $\sigma^2(\tau)$ in the observed variance $\sigma^2(\tau + \varepsilon)$. Item Response Theory (IRT) is a modern test theory framework which does not make an effort to single out a supposedly true score from the erroneous measurement, but instead pursues a probabilistic approach to quantify the characteristics and quality of items and tests. CTT and IRT are both introduced in more detail in Chapter 9.

5.3 CONCEPT INVENTORIES FOR MECHANICAL ENGINEERING

In this section, a (non-exhaustive) selection of CIs related to mechanical engineering are briefly presented. Because of the indisputable overlap of concepts used in engineering and physics, many of the instruments designed for physics education are also suitable for engineering education, and thus should not be left unmentioned. As the Concept Assessment Tool for Statics (CATS) is the central instrument in this dissertation, it is introduced in detail in Chapter 6.

THE FORCE CONCEPT INVENTORY (FCI) The development of CIs was initially sparked by the FCI (Hestenes et al., 1992) which is probably the most successful and most widely used CI in physics education research (Engelhardt, 2009). The success of the FCI in showing that student understanding of the force concept was much lower than expected also greatly promoted the discourse about physics education (Mazur, 1997; Garvin-Doxas et al., 2007; Engelhardt, 2009) and the development of CIs in other disciplines. Engineering education practitioners and researchers adopted the methodology and developed various CIs in the engineering disciplines (Montfort et al., 2009).

The focus of the FCI lies on the Newtonian force concept involving Newton's Laws, kinematics, the superposition principle and kinds of forces. The distractors were designed to address the misconceptions or "commonsense beliefs[2]" known to exist frequently among students (Halloun and Hestenes, 1985a). They were organized by Hestenes et al. (1992) in a taxonomy of six categories:

2 The term *misconception* is avoided by the authors and by many others to acknowledge the value of alternative conceptions.

- Kinematics, e. g. not discriminating velocity and acceleration
- Impetus, e. g. impetus is supplied by a "hit" and/or may dissipate along the way
- Active Force, e. g. only active agents exert forces
- Action/Reaction Pairs, e. g. greater mass implies greater force
- Concatenation of influences, e. g. last force to act determines motion
- Other influences on Motion, e. g. obstacles exert no forces

The impetus beliefs, for example, are associated with lack of understanding of Newton's First Law. Students often believe that moving objects received some kind of "impetus" that initially caused the motion. This impetus is "used up", which causes the object to slow down and stop, for example a ball being kicked once ("receiving impetus") will move and eventually roll to a stop when the "power", "energy", or "force" (as students usually call it) is used up. Other frequently observed beliefs are expressions of the conception of force as something active, e. g. only living beings or moving things exert forces on other objects they come in contact with, or motion implies the presence of an active force (and, consequently, no motion implies no presence of a force).

Because these beliefs are often reasonably grounded in everyday experience, the FCI "is not a test of intelligence; it is a probe of belief systems" (Hestenes et al., 1992, p. 142). Below, examples will be shown how some items allow to uniquely identify beliefs, while others do not: Common student beliefs on action/reaction force pairs involve the idea that one of the forces "wins" over the other, often the bigger or more active object. These beliefs are probed for example with item 11 (Figure 5) or with items 13 and 14 (Figure 6) that all address Newton's Third Law. For item 11, answer (E) is correct. Distractor (B) reflects the belief that only active agents (in this case the student pushing with his feet) exert forces. (A) and (C) are uncategorized. (D) may either indicate the belief that either the more active or the heavier agent exert the greater force. The design of item 11 does not allow to discriminate between these reasonings (and thus should be revised). With items 13 and 14, this discrimination is possible, because the heavier object (the truck) is not at the same time the more active one (the car). The correct answer is 13/14(A). Distractor (B) maps to the belief that the heavier object exerts the greater force, while students choosing (C) likely believe that the more active agent exerts the greater force. (D) is attractive to those students who believe that *only* active agents exert forces, and (E) to those who believe that obstacles exert no forces (which is similar to non-active agents).

* **Two students, student "a" who has a mass of 95 kg and student "b" who has a mass of 77 kg sit in identical office chairs facing each other. Student "a" places his bare feet on student "b's" knees, as shown below. Student "a" then suddenly pushes outward with has feet, causing both chairs to move.**

11. In this situation,

(A) neither student exerts a force on the other.
(B) student "a" exerts a force on "b", but "b" doesn't exert any force on "a".
(C) each student exerts a force on the other but "b" exerts the larger force.
(D) each student exerts a force on the other but "a" exerts the larger force.
(E) each student exerts the same amount of force on the other.

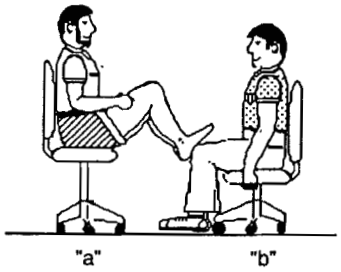


Figure 5: Item from the FCI addressing action/reaction pairs (Hestenes et al., 1992). As the active agent is at the same time the one with more mass, it is impossible to diagnose precisely the misconception underlying answer (D). The correct response is (E).

* **Refer to the following statement and diagram while answering the next two questions.**

A large truck breaks down out on the road and receives a push back into town by a small compact car.

13. While the car, still pushing the truck, is **speeding up** to get up to cruising speed;

(A) the amount of force of the car pushing against the truck is equal to that of the truck pushing back against the car.
(B) the amount of force of the car pushing against the truck is less than that of the truck pushing back against the car.
(C) the amount of force of the car pushing against the truck is greater than that of the truck pushing against the car.
(D) the car's engine is running so it applies a force as it pushes against the truck but the trucks engine is not running so it can't push back against the car, the truck is pushed forward simply because it is in the way of the car.
(E) neither the car nor the truck exert any force on the other, the truck is pushed forward simply because it is in the way of the car.

14. After the person in the car, while pushing the truck, reaches the cruising speed at which he/she wishes to continue to travel at a constant speed;

(same response options)

Figure 6: Another two items from the FCI addressing action/reaction pairs (Hestenes et al., 1992). The response options to both questions are identical (and reprinted for question 14 on the original test). Here, the lighter object is the active agent, which allows to discriminate between those students attributing the greater force to the active agent (C) and the ones attributing the greater force to the heavier object (B). The correct response is (A).

Student interviews were conducted to verify the reasoning behind the answer choices, revealing reproducible reasoning but also that false positive answers "were fairly common" (Hestenes et al., 1992, p. 148). This implies that the total score rather overestimates student understanding of the force concept. More formal establishment of validity and reliability was not conducted. The decision to omit this step was justified by Hestenes et al. (1992) with the similarities of the FCI to the Mechanics Baseline Test (MBT), which had been formally investigated before (Halloun and Hestenes, 1985b).

Unlike the CATS, for example, the FCI does not provide measurements for sub-scales of Newtonian thinking, despite the given taxonomy. Hestenes et al. (1992) recommend to use only the total score as a quantitative measure for individual students. Nevertheless, the individual responses students provide may be used on a qualitative level as an indicator which misconceptions must be addressed by instruction. It may thus be used as a diagnostic tool for this purpose. Furthermore, Hestenes et al. (1992) provide evidence that the FCI is also suitable for assessing the effectiveness of instruction, which has been done abundantly (e. g. Hake, 1998; Von Korff et al., 2016). Because it tests a belief system that can be changed and not some kind of disposition, the FCI is not suggested to be used as a placement test, while it may be a suitable tool to check student readiness for subsequent courses.

A total score of 80 % was initially regarded as a threshold score for diagnosing an expert-like Newtonian belief system (Hestenes et al., 1992)[3]. This decision is based on a correlation with the MBT (described below) which tests problem-solving skills of physics problems involving the basic Newtonian concepts. Only students with at least 80 % on the FCI were able to score also 80 % or higher on the MBT (Hestenes and Wells, 1992). Similarly, the data shown by Hestenes et al. (1992) suggest that 60 % are a second conceptual threshold that must be overcome for effective problem-solving as measured by the MBT.

Not all colleagues see a "threshold" in the data. This interpretation was criticized in an intense academic discourse on the central question of what the FCI actually measures (Huffman and Heller, 1995; Hestenes and Halloun, 1995; Heller and Huffman, 1995). A factor analysis conducted by Huffman and Heller (1995) on student FCI data did not reflect the taxonomy of commonsense beliefs proposed by the developers. This led the authors to the conclusion that the FCI does not measure a coherent force concept among students:

For details on factor analysis see Section 9.3.4.

> *"The items on the inventory appear to be only loosely related to each other, and instructors should be cautious about concluding that the inventory actually measures students' understanding of a 'force concept.' It seems more likely that the inventory actually mea-*

3 Hestenes and Halloun (1995) raise the threshold to 85 %

> *sures bits and pieces of students' knowledge that do not necessarily form a coherent force concept."*

Hestenes and Halloun (1995) argue in return that the factor analysis results presented by Huffman and Heller (1995) actually support their own claims because they show that the majority of students as non-Newtonian thinkers does not have a coherent conception of the force concept. Therefore, it cannot be expected that the factor analysis reflects the proposed taxonomy.

Despite this strong criticism, Huffman and Heller (1995) acknowledge that the FCI is still valuable as a diagnostic tool for evaluating instruction, which justifies the fact that it is still widely used, likely because there is no better CI available that addresses the Newtonian force concept in a similar fashion.

The FCI is described here in more detail than the following instruments because of its key role in the history of CIs as well as in the later parts of this dissertation.

THE MECHANICS BASELINE TEST (MBT) The developers of the MBT (Hestenes and Wells, 1992) state that it "tests the application of Newtonian concepts to simple kinematics and dynamics of a single particle." It is called "Baseline" because it covers only the very basic concepts of mechanics from introductory physics at any level (high school to university).

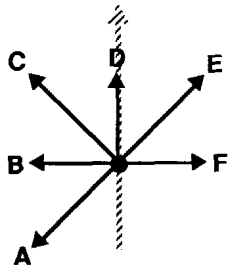
As a "sibling" of the FCI, the MBT also has its root in physics education. While the FCI targets students that have not been exposed to formal physics instruction and is therefore suitable as a pre-test, the MBT includes also items that require knowledge on quantitative problem-solving (see Figure 7). As such knowledge is expected only after formal physics instruction, the MBT is proposed only as a post-instruction test[4]. In this sense, the FCI and MBT complement each other. Correlating scores, Hestenes and Wells (1992) find that "[...] a good score on the [FCI] is a necessary but not a sufficient condition for a good score on the [MBT] or on other problem-solving tests on mechanics."

While the MBT may be a valuable instrument for measuring students' baseline understanding of mechanics in a physics curriculum, the strong focus on dynamics and kinematics makes it unsuitable as a post-test to a statics course.

THE ALTERNATIVE STATICS CONCEPT INVENTORY (ASCI) The CATS was formerly known as Statics Concept Inventory (SCI) and specifically designed for mechanical engineering curricula by Steif and Dantzler (2005). It is made up of 27 items that address 9 essential statics concepts such as drawing forces on separated bodies, equilibrium, friction or possible forces acting at different types of joints. The

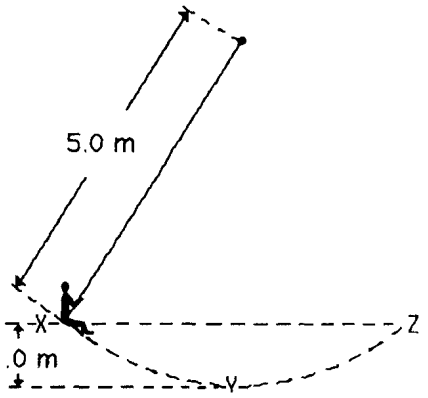
4 For advanced university courses it may be used as pre-instruction placement test.

19. The diagram at the right depicts a hockey puck moving across a **horizontal, frictionless** surface in the direction of the dashed arrow. A constant force **F**, shown in the diagram, is acting on the puck. For the puck to experience a net force **in the direction of the dashed arrow**, another force must be acting in which of the directions labeled **A, B, C, D, E**?



* **Refer to the diagram below when answering the next two questions.**

X and Z mark the highest and Y the lowest positions of a 50.0 kg boy swinging as illustrated in the diagram to the right.



11. What is the boy's speed at point Y?

(A) 2.5 m/s (B) 7.5 m/s
(C) 10. m/s (D) 12.5 m/s
(E) None of the above.

12. What is the tension in the rope at point Y?

(A) 250 N (B) 525 N (C) 7×10^2 N (D) 1.1×10^3 N
(E) None of the above.

Figure 7: Example of items on the MBT, taken from Hestenes and Wells (1992). Item 19 relates to an FCI item, while items 11 and 12 are much closer to quantitative end-of-chapter problems.

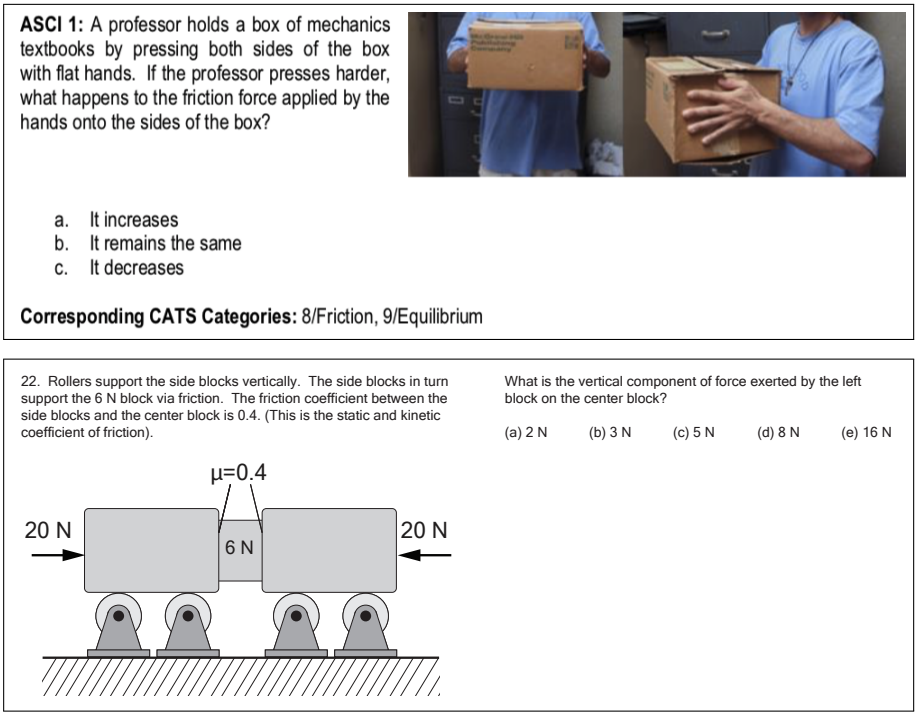


Figure 8: Example of an item on the ASCI, taken from Papadopoulos et al. (2016). This item is strongly related to CATS item 22.

CATS will be introduced in detail in Chapter 6. As the CATS developers stated that the instrument was not an informative measure for student understanding before they were exposed to statics instruction, Papadopoulos et al. (2016) developed the Alternative Statics Concept Inventory (ASCI) to be able to gauge student misconceptions at pre-instruction level. The development process was described as "not structured", but relying mainly on the teaching experience of the authors. Although it was not developed to map to the CATS concepts, all CATS concepts are represented (except for the Pin-in-slot concept) and the link is made explicit by the authors (see Figure 8).

The ASCI consists of ten items in a multiple-choice single-select format with the number of response options varying from two to five. Figure 8 shows an example of an item which seems to have been inspired by CATS item 22, but translating it into a more qualitative form. Many items are set in real life situations which are illustrated by photographs. However, technical terms such as "friction" and "torque" are used assuming that students are familiar with them, at least in an informal manner. Some items require skills beyond the scope of statics, such as estimating the slope of a road. These items often had unacceptable item statistics and should be revisited. Where applicable, the concept sub-scale correlations between the ASCI and the CATS were found to be reasonable, while the correlation of the totals scores on each test was moderate, as was correlation with exam scores.

The ASCI may have been an alternative to using the FCI as the pre-test instrument, had it been available at the time when data collection first began for the investigation discussed in Part III.

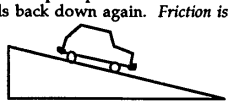
THE STRENGTH OF MATERIALS CONCEPT INVENTORY (SOMCI)
The first attempt to develop a Strength of Materials Concept Inventory (SOMCI) was not successful, possibly because it did not follow the systematic steps. Richardson et al. (2003) report bad item statistics but also systematic outline for a second attempt. However, personal communication with the second author indicates that, unfortunately, the project is currently not pursued any further.

THE FORCE AND MOTION CONCEPTUAL EVALUATION (FMCE)
The FMCE was developed by Thornton and Sokoloff (1998) to test student understanding of dynamics and kinematics. It puts a strong emphasis on graphical representations and the ability to transform between them, while also including items that rely only on natural language. It started out with 43 questions and evolved to 47 (Thornton et al., 2009) questions in total of which many relate to the same situations. Despite the strong similarities to the FCI, the focus of the FMCE is much narrower as it only covers forces and motion in one dimension. Furthermore, the format differs from typical multiple-choice tests in that is asks students to match repeatedly offered response options to the questions. For example, questions 8 to 10 (see Figure 9) refer to the three qualitatively distinguishable phases in a situation of a toy car rolling up a ramp and back down again (Phase 1: car is moving upwards, phase 2: car is at point of return, phase 3: car is moving downwards). In each question/phase, the students are asked for the direction of the net force on the car (which is down the ramp in every case).

Thornton and Sokoloff (1998) present evidence for the validity of the instrument such as strong correlations of student answers to items that probe the same concept as well as verification through open-ended responses that give insight into student reasoning.

Thornton et al. (2009) compared the FMCE and the FCI by administering both tests to every student as pre- and post-test. They found that both instruments were likewise able to measure higher gains in research-based interactive instructional settings compared to traditional instruction but each instrument may result in different evaluations depending on the student population or the particular instructional method. Because of its narrower focus, which is probed with more items, the FMCE is proposed as a more precise measurement tool for understanding of the Newtonian Laws than the FCI, which in turn probes for a broader understanding of the topic. The correlation coefficient of the totals scores is reported to be about 0.78.

8-10 refer to a toy car which is given a quick push so that it rolls up an inclined ramp. After it is released, it rolls up, reaches its highest point and rolls back down again. *Friction is so small it can be ignored.*



Use one of the following choices (A through G) to indicate the **net force** acting on the car for each of the cases described below. Answer choice **J** if you think that none is correct.

(A) Net **constant** force down ramp
(B) Net **increasing** force down ramp
(C) Net **decreasing** force down ramp
(D) Net force zero
(E) Net **constant** force up ramp
(F) Net **increasing** force up ramp
(G) Net **decreasing** force up ramp

_____8. The car is moving up the ramp after it is released.
_____9. The car is at its highest point.
_____10. The car is moving down the ramp.

Figure 9: Example of an item on the FMCE, taken from Thornton and Sokoloff (1998).

THE R-FCI Nieminen et al. (2010) developed a representational (R-) variant of the FCI. For nine original FCI items, they created two additional isomorphic representations of the multiple-choice response options, resulting in a 27-item instrument that was tested in a Finnish high school context.

For many CIs, the development did not follow a systematic approach as the one described in the previous section. Lindell et al. (2006) compared several validity studies of CIs and showed that they varied in multiple aspects such as which statistics are reported or which type of validity aspect was addressed (if any). Adopters of existing CIs should thus select the instrument with care by consulting the validity studies associated with the instruments as the mere existence of an instrument is not a guarantee for its quality (see also differences in CI quality found by Jorion et al., 2015). Furthermore, the adopter must ensure that the instrument is suitable to use in the context of the adopter's research (Lindell and Ding, 2013), a condition which motivated the revalidation study presented in Part I of this dissertation.

In Pellegrino et al. (2013), the CATS was selected as an illustrative example for the design and interpretation of concept inventories in engineering education. The instrument is described in detail in the following chapter.

6

THE CONCEPT ASSESSMENT TOOL FOR STATICS (CATS)

This chapter starts with background information on the development of the CATS and its key characteristics, followed by a description of the cognitive steps necessary to respond correctly to the items. Finally, literature on the investigation of validity and on research using the instrument as a measurement is reviewed.

The Concept Assessment Tool for Statics (CATS) was developed with focus on the concepts required for the analysis of multiple, connected bodies. It was first published as the Statics Concept Inventory (SCI) in 2005 (Steif and Dantzler, 2005) and has been revised through several versions (Steif and Hansen, 2007). While the latest is version 6, all investigations in this thesis relate to version 4, the current version at the time of beginning data collection for the main investigation presented in Part III.

See Appendix A for CATS version 4.

Inspired by the FCI and the work in PER, Steif (2004) brought forward a suggestion of formulating essential concepts and skills in Statics to open a discussion on the necessity and phrasing of such an articulation. According to Steif (2004), the main distinction between skills and concepts is that skills can be mastered by rote learning. Apart from the mathematical skills required for solving Statics problems, the following four skills were articulated as essential:

"S1. Discern separate parts of an assembly and where each connects the others"

"S2. Discern the surfaces of contact between connected parts and/or the relative motions that are permitted between two connected parts"

"S3. Group separate parts of an assembly in various ways and discern external parts that contact a chosen group"

"S4. Translate the forces and couples which could be exerted at a connection (e.g., there is only a force in a known direction) into the variables, constants, and vectors that represent them"

The four essential concept clusters were later described as:

"C1. Forces are always in equal and opposite pairs acting *between* [original emphasis] bodies, which are usually in contact."

"C2. Distinctions must be drawn between a force, a moment due to a force about a point, and a couple. Two combinations of forces

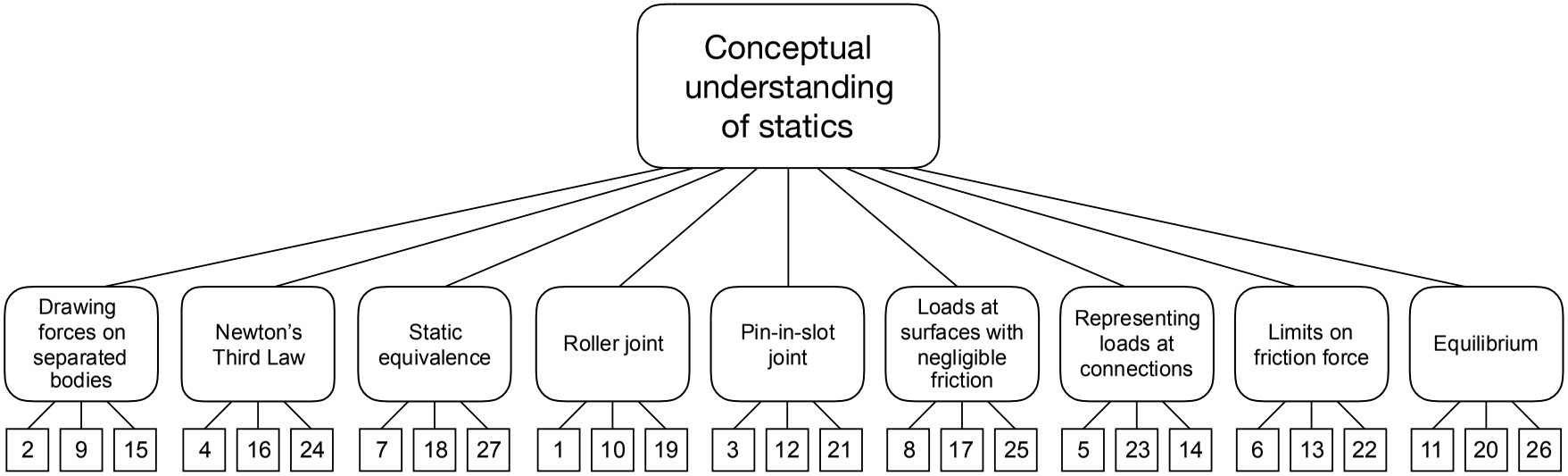


Figure 10: Model of the theoretical underlying structure of the CATS. The entire test, composed of nine concepts with three items each, is supposed to measure the single construct 'conceptual understanding of statics'.

and couples are statically equivalent to one another if they have the same net force and moment."

"C3. The possibilities of forces between bodies that are connected to, or contact, one another can be reduced by virtue of the bodies themselves, the geometry of the connection and/or assumptions on friction."

"C4. Equilibrium conditions always pertain to the external forces acting directly on a chosen body, and a body is in equilibrium if the summation of forces on it is zero and the summation of moments on it is zero."

(Steif and Dantzler, 2005)

The CATS was developed based on these concepts and the following, frequently observed student errors:

1. "Failure to be clear as to which body is being considered for equilibrium."
2. "Failure to take advantage of the options of treating a collection of parts as a single body, dismembering a system into individual parts, or dividing a part into two."
3. "Leaving a force off the free body diagram (FBD) when it should be acting."
4. "Drawing a force as acting on the body in the FBD, even though that force is exerted by a part which is also included in the FBD."
5. "Drawing a force as acting on the body of the FBD, even though that force does not act directly on the body."
6. "Failing to account for the mutual (equal and opposite) nature of forces between connected bodies that are separated for analysis."
7. "Ignoring a couple that could act between two bodies or falsely presuming its presence."
8. "Not allowing for the full range of possible forces between connected bodies, or not sufficiently restricting the possible forces."
9. "Presuming a friction force is at the slipping limit (μN), even though equilibrium is maintained with a friction force of lesser magnitude."
10. "Failure to impose balance of forces in all directions and moments about all axes."
11. "Having a couple contribute to a force summation or improperly accounting for a couple in a moment summation."

(Steif and Dantzler, 2005)

Prior to its development, Steif (2004) argued that an instrument for measuring conceptual understanding like the FCI for Statics would need to address all the essential concepts without requiring the mentioned skills, and concludes that the mechanical systems on such a test would have to be sufficiently simple to not allow errors due to missing skills, but only based on conceptual lapses. Subsequently, Steif and Dantzler (2005) introduce the first version of the Statics Concept Inventory (SCI), as the CATS was named initially. This first version was already composed of 27 items, but only eight concept categories with a variable number of items associated with each concept. The concept category Newton's Third Law was introduced in a later version. The structure of the CATS evolved over time to nine different concepts:

1. *Drawing forces on separated bodies* (also sometimes referred to as *Free-body diagrams* or *FBD*)
2. *Newton's Third Law*
3. *Static equivalence*
4. *Roller joints* (also sometimes referred to as *Roller(s)*)
5. *Pin-in-slot joints* (also sometimes referred to as *Pin-in-slot* or *Slot*)
6. *Loads at surfaces with negligible friction* (also sometimes referred to as *Negligible friction* or *Frictionless contact*)
7. *Representing loads at connections* (also sometimes referred to as *Representations*)
8. *Limits on friction force* (also sometimes referred to as *Limit of friction* or *Friction*)
9. *Equilibrium*

. Understanding of these concepts is tested by three items each in form of multiple-choice questions, evenly distributed over the 27-item test instrument. Based on these nine concepts, the CATS is supposed to measure one higher-order single construct, i. e. conceptual understanding of statics. Therefore, the scores on all items may be added to form a single test score. Figure 10 illustrates the theoretical structure of the instrument. The item numbers indicate their sequence on the administered test.[1]

1 It should be noted that in other literature on the CATS, the authors most often refer to the item sequence ordered by concept (as ordered in Figure 10 from left to right).

PURPOSE AND INTERPRETATION As mentioned above, the concept of validity requires information on the purpose of a test, including instructions on the interpretation of test results as well as formal administration guidelines. Ideally, both is provided by the developers. The CATS developers state that "[t]he underlying theoretical construct of this instrument is 'statics conceptual knowledge.'" (Steif and Dantzler, 2005, p. 369). A more detailed interpretation guide with respect to the score range is not given. As mentioned in Chapter 5, CIs in general are applicable to evaluate the effectiveness of instruction, as well as to formatively assess individual students, and to diagnose common misconceptions and their persistence after instruction. The developers explicitly address the latter and claim that "[o]n the basis of [the CATS], one can infer which concepts students in general tend to have the most difficulties with, as well as the misconceptions that appear to be most prevalent" (Steif and Dantzler, 2005, p. 371). While it is not explicitly stated that the purpose of the CATS is assessment of instruction, the statement that "assessment of conceptual understanding can help instructors to gauge the effectiveness of new teaching methods and approaches" (Steif and Dantzler, 2005, p. 363) indicates that the authors intend it to be used for assessment of instruction as well. The main purpose of the CATS is thus *identification of student misconceptions* to inform instruction and misconception research, while the *assessment of instruction* poses the secondary purpose. The proposed time limit is 60 minutes, which was found to be "plenty of time in which to complete this test" (Steif and Hansen, 2007).

While the CATS was originally intended to work as both, pre-test and post-test (Steif and Hansen, 2006a), its validity as pre-test has been questioned (Steif and Hansen, 2007).

6.1 COGNITIVE STEPS TO RESPOND CORRECTLY TO THE CATS ITEMS

In this section, a line of argumentation to arrive at the correct response will be laid out for each item of the CATS. In addition, the reasoning behind the distractors is explained. The items are grouped by concept.

Q02, Q09, Q15 - DRAWING FORCES These items are designed to measure whether the correct forces are included in a FBD. They address student errors 1-5. An essential rule for drawing free body diagrams is that only those forces are drawn that are directly exerted *by other* bodies *on the* body or the collection of bodies. For example, for item Q02, only the cord forces T_B, T_C, T_E, and T_F as well as the weight forces of the included bodies, W_2 and W_3, must be drawn. Inner forces (here T_D) or weight forces on external bodies (e. g. W_5) must not appear. Similar reasoning applies to Q09 and Q15, though the for-

mer case includes normal contact forces instead of tension forces by cords.

The distractors address the following misconceptions, some of which occur in multiple distractors (Steif and Hansen, 2007):

- IntF: Internal forces are included.
- WnotF: Contact forces are labelled as weight forces of external bodies, (acknowledging that they have the same magnitude and direction, but ignoring that they are different types of force and acting on different bodies).
- W+F: The weight force of external bodies is included in addition to the contact force transmitting the effect of this weight.

Q04, Q16, Q24 - NEWTON'S THIRD LAW These items are designed to measure the understanding of the principle that "action equals reaction" (C1, student error 6). All three items show a frame subjected to arbitrary external forces. The frame is composed of several parts which are connected by a frictionless pins, i. e. moments do not need to be considered. Each pin exerts a force on each of the two connected frame parts. The items ask for the directions of these forces. The direction of any single one of these forces cannot be determined because of the arbitrary external forces. Instead, the interdependence of the directions of the forces has to be recognized. The pin is in static equilibrium. By Newton's second law, this means that $\Sigma\vec{F} = 0$. The two forces acting on the pin by the two frame parts must induce this state, i. e. they must be equal and opposite. By Newton's third law, the forces by the pin on the frame parts must be equal and opposite as well. While all response options seem to show forces equal in magnitude, there is only one option with opposite forces in each of the three items.

The distractors address the misconceptions that

- 2Force: The forces in the bars must act parallel to the bars like in a rod/two-force body.
- Shear: The forces in the bars must act perpendicular to the bars like a shear force.
- Perp: The forces must act perpendicular to each other like independent force components.
- SameDir: The forces must act in the same direction.

(Q04 has slightly different distractors than Q16 and Q24, probably because the frame parts are connected at a right angle, which would cause some distractors to be addressed twice. For example, the distractor which addresses a combination of Shear and 2Force in Q16 and Q24 would be identical to Perp.)

Q07, Q18, Q27 - STATIC EQUIVALENCE These items ask for replacing a load or a set of loads acting on a body by another set of forces and/or moments, so that the body is still in static equilibrium (C2 and C4, student errors 10 and 11). In general, there are infinitely many possibilities to replace a set of loads by another one while maintaining equilibrium. These items must hence be solved by exclusion of the distractors. An adequate replacement must have the same net force and net couple as the load(s) to be replaced. For items Q07 and Q18, only a single load is to be replaced, a moment in one case and a force in the other. Item Q27 is more complex because two forces must be considered. Here, the larger force (10 N) can be split up into an upward force of 7 N to form a couple with the force acting downward at point A, and a single upward force of 3 N at point B. Considering the distance between points A and B, the equivalent loads replacing the drawn forces must have a net couple of 280 N-mm and an upward net force of 3 N in order to maintain equilibrium with the other forces.

The distractors are diverse, because of the different loads to be replaced, but the following misconceptions can be found in all items:

- M=F: A single force can be equivalent to a moment as long as $M = F \cdot d$.
- M=Md: a load can replace a load of the same kind (single force or moment) acting elsewhere, and the change in point of application is accounted for by changing the magnitude.

Another misconception addressed by Q07 is that

- CentrM: A force couple can replace a moment if it is centered about the same point as the moment. The distance between the forces is ignored (Q07 (d)).

Q01, Q10, Q19 - ROLLER JOINT These items are designed to measure understanding of forces that can be exerted by a certain support or joint, in this case the roller (C3, student errors 7-8). Between the roller and the respective body (block, platform or L-shaped arm) in contact, there is friction. Based on this information only, the roller could exert both, a tangentially and a radially oriented force on the body. However, by Newton's Third Law, any tangential force exerted by the roller on the body must be equally exerted by the body on the roller. This force on the roller would result in a moment about the pin supporting the roller. Since the pins are frictionless and no other forces on the roller are exerting a moment, the sum of moments would be non-zero for the roller. Therefore, the roller cannot exert a tangential force on the body in static equilibrium, it can only exert normal forces. (Note that one must assume the block in Q01 and the platform in Q10 to be subjected to a weight force acting downward although none is indicated. This was found to be an error but it does not seem to be noted by or confusing to the students.)

The distractors on the Roller concept address the following misconceptions:

- 2Force: The force acts parallel to the member on which it acts like in a two-force member.
- ApplF: The direction of the applied external force dictates the direction of the force at the connection.
- arctan(μ): Another common distractor to all Roller items is that the force exerted by the roller acts at an angle of arctan(μ) (which is the maximum angle of an incline before objects start sliding down due to the limit on the static friction force). It is unclear whether this distractor reflects a real misconception or whether it was simply added as a convenient fifth response option.

Q03, Q12, Q21 - PIN-IN-SLOT JOINT The Slot items target the same concept and errors as the Roller items (C3, student errors 7-8), but the joint is a different type. The pin in the slot is frictionless. Therefore, there cannot be a force on the pin in the direction along the slot, any force must be perpendicular to the slot. A moment can also not be exerted because there is neither friction nor a form fit. Using simple geometric reasoning, the direction of the force can be found. (Note that in Q03, depending on the direction of the force by the spring, the orientation of the force is not uniquely defined, but there is only one response option that follows the logic laid out above.)

The distractors on the Slot concept address the following misconceptions (Steif and Dantzler, 2005):

- 2Force: (see Roller concept)
- Motion: The force acts in the direction of the slot, the direction in which motion is unhindered.
- Moment: A moment is induced by the applied force.
- ApplF: (see Roller concept)

Q08, Q17, Q25 - FRICTIONLESS CONTACT Another type of contact which reduces the possibilities of forces are contacting surfaces with negligible friction. The items are designed to measure whether students understand the consequences of negligible friction for the direction of possible forces (C3, student errors 7-8). Since the contact between the surfaces is frictionless, there can be no force component along the surface but only normal to it. Moments can also not be transmitted as this would only be possible by form fit. None of the cases given in Q08 and Q17 are possible, because they do not assume pure normal forces. The pin in Q25 can exert forces in all directions in the plane, but it cannot exert a moment.

In Q08, the previously discussed misconceptions 2Force (force parallel to arm) and ApplF (applied moment causes force perpendicular to arm) are also at play. The wording of the response options is the same for all three items: two situations must be assessed for whether the depicted reactions are possible. As there are only four possibilities of combining the assessments, a "not enough information" option (e) was added in order not to deviate from the five-choice structure.

Q05, Q23, Q14 - REPRESENTATION OF LOADS AT CONNECTIONS The three items address three different types of connections - a frictionless pin, a welded joint, and a cable - with regard to possible reaction forces to arbitrary external loads (C3, student errors 7-8).

In Q05, the support is a pin. The pin is frictionless, hence no moment can be exerted by the pin on the plate. In terms of forces, the general case must be selected because the forces applied to the plate are arbitrary. The most appropriate representation thus has two degrees of freedom: the magnitudes of a horizontal and a vertical force component.

In Q23, the rigid (welded) connection of the two parts allows for all three possible loads in two dimensions (F_h, F_v, and M) to be exerted by one part on the other. As the external loads are arbitrary, it is most appropriate to assume three independent loads at the connection.

Moments cannot be exerted by cables, which eliminates distractors (b) and (d) in Q14. The direction of the force is given by the cable in that it can only exert tension forces along its own direction. This allows to eliminate distractors (c), and (e) which describe more general cases. Option (a) is therefore most appropriate.

Q06, Q13, Q22 - LIMIT OF FRICTION The items are designed to address student understanding of static friction and its character as a reaction force (C4, student errors 9-10). According to the instructions, all bodies are in static equilibrium. For Q06 and Q13, to achieve equilibrium in the horizontal direction, the unknown friction force must be equal and opposite to the only external horizontal load (Q06: 8 N, Q13: 10 N). For all the friction items, one may choose to verify that the limit of the static friction force $\mu \cdot F_N$ is larger than the force required to achieve equilibrium, but it is not necessary if the instructions are trusted. If the limit were exceeded there would be acceleration and thus a conflict with the instructions.

The static friction case is implied more strongly in Q22 through the item description compared to the other friction items ("The side blocks *support* the 6 N block via friction" (emphasis added). Due to symmetry, the friction force on the right side of the block will be equal to the friction force on the left side of the block. Applying the equilibrium condition for the vertical forces on the middle block results in a required friction force of 3 N on either side.

Typical misconceptions addressed by the Friction concept are (Steif and Hansen, 2007):

- MuN: The product of friction coefficient and normal force (μN) is not seen as the limit but as the actual friction force.
- F-MuN: The tangential force is the driving force minus "what friction (again μN) takes away". Here, the idea of friction as some kind of dissipation of a force may play out.

These misconceptions appear multiple times in the distractors. Variants include the use of an incorrect value for the normal force and variation in the direction of the force.

Q11, Q20, Q26 - EQUILIBRIUM These items are designed to measure student understanding of the equilibrium conditions (C4, student errors 10-11). The general approach to these items is to check the three equilibrium conditions in two dimensions: $\Sigma F_h = 0$, $\Sigma F_v = 0$, and $\Sigma M = 0$.

In Q11, the hand represents a fixed support, which could exert forces and moments in all directions, if necessary. In this specific case, the external load is known. The equilibrium conditions for forces and moments show that a horizontal force and a counter-clockwise moment must be exerted by the hand.

Q20 shows two load cases, for which the possibility of fulfilling the equilibrium conditions for forces and moments must be checked qualitatively. In case (I), independently of the magnitude of the forces shown, there is a non-zero vertical force: $\Sigma F_v \neq 0$, which not only results in a non-zero net force, but inevitably also in a non-zero net moment. In case (II), the lines of action of the three forces do not intersect in a single point and the resultant moment is non-zero, $\Sigma M \neq 0$. Consequently, neither of the cases allows for static equilibrium.

Evaluating the equilibrium conditions in the depicted load case in Q26 yields $\Sigma F_v \neq 0$, while equilibrium can be fulfilled for the horizontal forces and the moments. The additional load must hence be a vertical force *and* its point of application must be chosen such that the lines of action of all three forces intersect in one point. Only then can an additional force be applied without introducing a net moment. These conditions only allow an upward force acting at P (or R, but that is not available as an option).

The distractors express a variety of cases where one or more of the equilibrium conditions are violated. The misconception that a moment can be equivalent to a single force (M=F) plays out here as well (e. g. Q11(b)). Another possible misconception is that it is sufficient to check $\Sigma M = 0$ for one point. This rule holds as long as $\Sigma F_h = 0$ and $\Sigma F_v = 0$ are true. When there is an imbalance in terms of forces, $\Sigma M = 0$ cannot be fulfilled globally. In Q11(a), for example, the moments balance about the point of intersection of the lines of

action of the forces, but this is the only point for which this holds. The unbalanced vertical force component impedes the possibility of a global balance of moments. The expected response pattern for students with this misconception is Q11(a)-Q20(c)-Q26(b) (although in Q26, the part would be seen as being already balanced about point P). (Note that for generating this response pattern, the imbalance of forces would have to be systematically ignored by the student, which makes it more unlikely to diagnose this misconception.) Distractor (e) is the best representation of all reactions that the hand as a fixed support could exert, if necessary: a vertical and a horizontal force component and a moment. Students selecting this response probably struggle with the concept that the forces and moments by a support are reactive and determined by the applied loads.

6.2 LITERATURE REVIEW

The literature on the CATS can be divided into research on the instrument itself and on research using it as a measurement tool. Examples from both those categories are given in the following sections.

6.2.1 *Prior validity research on the CATS*

Upon introducing the CATS, Steif and Dantzler (2005) presented first psychometric evidence based on a sample size of 245 students to assess the reliability and validity of the results obtained with the instrument. An analysis of variance (ANOVA) showed no effects of race or gender. The item analysis indicated that all but one item were in the acceptable ranges of difficulty and discrimination. Reliability was assessed as strong by means of Cronbach's α ($\alpha = 0.89$). Content validity was established through inspection by experienced instructors as experts as well as inspection of written student answers to statics problems that reflect the typical errors. Criterion validity was established by correlation with exam performance (Spearman's rho $= 0.547$), due to lack of a better comparable instrument. A confirmatory factor analysis (CFA) was performed to confirm the proposed structure of only eight factors at the time for construct validity. The results are described as merely acceptable and likely motivated the changes that led to the restructured CATS version with nine concept categories.

Item 20[2] was reported by Steif and Dantzler (2005) to be the most difficult and weakly discriminating item. Newcomer and Steif (2008) state the assumption that the weak performance on item 20 is partially caused by a mismatch to the style of problems students are used to (*"There is the tacit assumption in typical problems that equilibrium is possible."*), and they criticize typical statics instruction in that it does

2 item 26 in concept order

not prepare students to assess the possibility of equilibrium. They suggest that *"[i]nstruction that promotes the contemplation of the possibility of equilibrium, rather than only the quantitative implications of its presence, may both improve performance on this assessment task and may have broader benefits in preparing students for using mechanics in engineering practice."* They conducted an analysis of answers including written explanations to the item in order to gain insight into the considerations made by the students, how these considerations relate to the selected answer, and how thinking evolves with instruction. The written explanations were coded with respect to whether an equilibrium condition (force or moment) was applied "(1) never, (2) insufficiently, (3) appropriately, just as needed, or (4) in all cases" (Newcomer and Steif, 2008, p. 488), and evidence for the consideration that forces influence moment equilibrium was monitored.

By correlation analysis, it was shown that the selected answer on the multiple-choice item tended to match the coding of the explanations, e. g. students who selected (c) tended to produce codes (1) or (2) on force equilibrium, and codes (3) or (4) on moment equilibrium. These results indicate that the item validly measures whether students are able to *consistently* apply both force and moment equilibrium conditions.

Steif and Hansen (2006a) analyzed CATS data[3] from ten engineering mechanics courses held at seven different institutions. Cronbach's α was reported as 0.82, which is evaluated as a good measure for reliability and internal consistency. Furthermore, discrimination indices are reported as well as Pearson's r correlation with course examinations, where data was available. Exam correlations range from about 0.2 to about 0.6 in both, Statics and Mechanics of Materials courses. These values are seen as "quite meaningful", "[g]iven that the inventory asks questions on highly isolated and idealized aspects of statics". Correlations among multiple examinations within each class (midterms and final) were of the same magnitude.

In addition to evaluating the total score correlations, the success in performing specific analytical steps required on individual exam problems were related to the success on the respective CATS concepts. Students that erred on a specific analytical step on the exam (e. g. analyzing forces exerted by a roller) also had a significantly lower sub-score on the related CATS concept than students who did not err on this step. Results also showed that the CATS has predictive power of such kind even on content that is not specifically addressed by the CATS, like determining continuous shear forces and bending moments. This supports the claim that the CATS concepts are central even though they are not all-encompassing.

Evidence for the power of the CATS as a misconception diagnostic tool was reported for example by Steif and Hansen (2006a) and

3 collected with a prior version of the test than the one investigated in this dissertation

Steif and Hansen (2007). The latter showed that patterns which differ from random guessing can be seen in the choice of distractors and present more psychometric analyses of a slightly modified version of the CATS on a larger data set than the studies before. Changing the administration mode from paper and pencil to online provided additional insights into the time required for completing the test. It was found that "[t]here is no noticeable pattern in the variation of mean scores for [...] examinees with time above 25 minutes".

They present various methods to analyze different aspects of the data. For example, to avoid the problem that correlations between items on the same concept are low if their difficulties differ strongly, they used an alternative approach to show that the items test the same concept despite low correlation: Students who fail on the easier item are also more likely to fail on the more difficult item, and students who answer the more difficult item correctly are also more likely to answer the simpler item correctly.

While the CATS had been used previously as a pre-test (Steif and Hansen, 2006a), it was re-evaluated here to be of little value for this purpose. Comparing the actual score distribution to the expected distribution in case of random guessing showed strong similarities, and - unlike post-test scores - pre-test scores were almost uncorrelated with exam scores, and thus have no predictive value (with the exception of high scorers).

Not only the developers themselves report evidence for validity. For example, Anderson et al. (2009) performed a correlation analysis among all the concepts. For the most part, they did not find significant correlations between the concepts, but between the individual concepts (except for Static Equivalence) and the total score. A linear regression analysis furthermore revealed that the course grade is most sensitive towards performance on the Equilibrium concept.

While these results were more of a side product of the study, Jorion et al. (2015) performed an extensive validation of the CATS. They proposed an analytical framework using complementary psychometric analyses to validate the generic claims that CIs can measure student understanding of the top-level construct, as well as at the concept level, while diagnosing typical misconceptions. The analyses encompass CTT, IRT, and exploratory factor analysis (EFA) followed by an optional CFA or diagnostic classification modeling (DCM). For illustration purposes, they applied their proposed framework to three instruments, including the CATS. The claims that the CATS can measure student understanding of statics at the top-level construct as well as at the concept level could be confirmed. Evidence for the instrument's diagnostic power as a misconception detector, however, was not strong.

This framework is partially applied in the following chapters for the revalidation of the CATS and is introduced in more detail in Chapter 9.

The CATS sample used by Jorion et al. (2015) consisted of 1372 students from various courses and institutions collected online during

one academic year. In the CTT framework, most item difficulty indices ranged from 0.25 to 0.78 and most item discrimination indices ranged from 0.20 to 0.65. The only exception is again item 20 (item 26 in concept order), which is extremely difficult and poorly discriminating. As a measure for reliability, Cronbach's α of the total score was reported to be 0.84, which was interpreted as "good reliability for an assessment used for low-stakes purposes". Items 10, 14 and 20[4] were identified to not contribute to the internal consistency of the instrument, which suggests the need for revising these items. The precision of the measurement based on CTT using total score descriptive statistics (mean and standard deviation) and Cronbach's α was calculated as 12.8 ± 2.02, meaning that, due to measurement error, the true scores of individual students scoring between 11 and 15 points may not differ.

Parallel to CTT analysis, the CATS data was analyzed with IRT methods. A two-parameter logistic (2PL) model fit the data best, indicating no strong influence of students guessing the correct answer. Model-to-data fit was evaluated to be close for nearly all items, although it is not clear from the paper by which standards. Only items 20 and 10[5] perform poorly in the fit. Those items were noted before in the CTT analysis.

For the structural analysis, only item 20 was removed because item 10 still correlated well with the other items in its concept cluster. The EFA suggested a structure close to the one intended by the developers with some deviations. Instead of the nine-concept structure, eight factors were suggested by a parallel analysis. The Equilibrium concept cluster did not emerge in the EFA, likely because one *Equilibrium* item was removed prior to the analysis. In addition, the *Representation* items and the *NeglFric* items did not consistently load onto the same respective factors. Jorion et al. (2015) conclude "that the developer's original categories account for student performance on most of the items", but that in case of the three deviating concept categories "students may be using complex, overlapping conceptual understandings to answer the items". A CFA was conducted for two models, one assuming independence of the concepts, and one including the higher order factor of statics conceptual understanding (see Figure 10). Unlike the independence model, the higher order model was found to be well fit to the data, confirming that the concept categories are linked by a top-level factor which can be assumed to be overall understanding of statics. The detailed results are a useful reference for the revalidation study in Part I.

4 items 11, 21, and 26 in concept order

5 items 26 and 11 in concept order

6.2.2 *Research using the CATS as a measurement instrument*

After the CATS had been established as a valid measurement instrument for statics understanding, it has been used in many studies for assessing the effect of various instructional interventions, for example, hands-on experiments (Coller, 2008), writing assignments (Venters et al., 2014), online homework (Arora et al., 2013), or short conceptual worksheets followed by discussion of misconceptions (Steele et al., 2014). Others used the CATS for research on student cognitive skills independent from instructional strategies. For example, Litzinger et al. (2008) investigated student problem-solving skills and behavior in the initial steps of drawing FBDs and setting up equilibrium equations. The students had previously been categorized into "strong" and "weak" students as a result of multiple measures, including CATS. As one would expect, "weak" students had more difficulties in drawing correct FBDs and setting up the equilibrium equations than "strong" students. Only in terms of invoking the condition for moment equilibrium, they were equally challenged. This result aligns with the findings by Newcomer and Steif (2008) laid out above. Based on the total CATS scores and the evaluation of the student problem-solving skills, the authors conclude that good conceptual understanding is not sufficient for successful problem-solving, because difficulties remain in applying the conceptual knowledge to typical statics problems. This statement is certainly plausible, but this does not mean that conceptual understanding is not beneficial to problem-solving skills.

One point to be criticized is that it is not clear whether the authors analyzed the CATS data on the concept level. The two problems given to the students required conceptual understanding associated with only four of the nine CATS concept categories: Drawing forces, Representations, Negligible friction, and Equilibrium. The maximum total CATS score in the sample was only 13 out of 27, indicating remaining conceptual problems, and the total score does not reflect whether the students had understanding of these concepts or rather of the ones irrelevant to the given problems.

The CATS was furthermore used to explore differences and similarities between the educational context and the work field. Using the same student data set as Jorion et al. (2015), Brown et al. (2019) compared student performance on the CATS to that of 95 civil engineering practitioners. They found that students outperform practitioners in terms of total score and that the difference in performance mostly attributes to four concept categories, namely Newton's Third Law, Static equivalence, Pin-in-slot joints, and Negligible friction, while the performance on the remaining concepts was statistically comparable. A comparison on the single item level revealed that these results do not necessarily apply consistently to every item of a concept. The authors offer three possible interpretations: (1) The use of concepts

depends on context in the sense of *situated cognition* and the setting in the CATS items causes a "contextual disconnect" (Brown et al., 2019, p. 131). Practitioners are focused on complex problem-solving and use concepts as a tool while for students, learning the concepts is the goal itself. (2) Practitioners still have misconceptions despite their practical experience[6]. (3) Unfamiliarity with the representation of concepts can hinder performance. They do not hypothesize that the weak performance on the four concepts may furthermore indicate that these concepts are less essential in the work environment than expected, which may also be a plausible explanation.

Differences between practitioners and students also showed in a prior factor analysis suggesting a four-factor structure for practitioners (Ha et al., 2017) contrary to the eight-factor structure found for students (Jorion et al., 2015), or the hypothesized nine-factor structure suggested by the developers. In an attempt to explain these factors, the authors tentatively suggest that the items may factor according to the perceived shapes of the bodies as 3D blocks, 2D planes, and 1D rods or bars.

6 This phenomenon has been documented in the higher education setting among pre-college teachers and physics graduate students (e. g. Shaffer and McDermott, 2005)

Part I

REVALIDATION OF THE CONCEPT ASSESSMENT TOOL FOR STATICS IN THE GERMAN CONTEXT

> *"Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use."*
>
> (American Educational Research Association et al., 2014, p. 14)

7

MOTIVATING THE REVALIDATION STUDY

For motivation, this part starts with a fictitious story about Sam, a young astronaut and researcher. Sam was told that astronauts quickly loose muscular mass on their missions. She wanted to find out more about the development of the loss rate over the course of a mission by keeping track of her own body mass on her next mission to the moon. To obtain reliable data, she decided to use the very same instrument for every single measurement. She brought her personal scale which, from her experience, always produced reliable measurements. She took the first measurement before the launch. The second measurement was taken just minutes after touch down on the moon. She was surprised to find out that she had gone from 70 kg before the launch to under 12 kg after landing. Further measurements on the moon during the following days indicated no more significant reduction, so she assumed saturation and stopped the measurements. The final interpretation of the data was that the loss rate is highest during the initial travel phase and nearly zero thereafter.

Obviously, Sam missed an important aspect: The validity of the interpretation of the scale's measurements had only been established for use on earth. The interpretation of measurements taken outside of this context led to false conclusions. In this case, the output of the selected instrument is interpreted as a mass, while the instrument actually measures weight and has been calibrated for use on earth. The change in location had an effect on the weight force and thus on the output as mass. The difference in local gravitational acceleration on the moon and on earth must be considered for a physically correct interpretation.

The measurement instruments used in engineering education research are rarely body weight scales, but usually tests such as standardized diagnostic instruments or concept inventories as introduced in 5. Nevertheless, a possible failure of validity due to context dependency equally applies. As "[v]alidation is the joint responsibility of the test developer and the test user", (American Educational Research Association et al., 2014, p. 13), test users must ensure that their interpretations of the test data are valid, under consideration of possibly different purpose or context of application. In this sense, a revalidation of instruments under consideration of the specific study context and design is essential (Lindell and Ding, 2013).

To answer the first research question of this dissertation, an already existing instrument is applied in a different national context than the one in which it was developed. Using a CI in a different national con-

text often requires adaptations which should be carefully validated. For example, Benegas and Flores (2014) used a Spanish translation of the "Determining and Interpreting Resistive Electric Circuits Concepts Test" (DIRECT) (Engelhardt and Beichner, 2004) that was validated by experts in the field. Mazak et al. (2014) developed a Spanish version of the CATS and investigated the effect of the translation on performance in a small-scale control-vs.-experimental-group study. Differences due to national context are not limited to language: Incompatibilities in terms of notations, conventions or course content may also affect the validity of the test result interpretation. Even for identical course content, different learning goals might be pursued. Last but not least, distractors might be culture-dependent. The revalidation process must thus go beyond language and translation issues.

The pursued revalidation process is lengthy and costly because it requires to collect and analyze various types of data. Still, it is less lengthy and costly compared to developing a new instrument because the validation process is a subset of the development process. In case the result of the revalidation study suggests that available instruments should not be used in the new context, a new instrument must be developed.

The purpose of this part is to present pieces of evidence for the validity of CATS in Germany. It addresses the following research question:

RESEARCH QUESTION Is the Concept Assessment Tool for Statics (CATS) a suitable measurement instrument for the investigated context of an introductory engineering mechanics course in the German higher education system?

This research question can be further specified as follows:

If the German CATS version is administered in the German higher education context:

1. Is the total score a valid interpretation of the level of conceptual understanding of statics?

2. Do the proposed concept subscales provide valid and reliable information on student understanding of those concepts?

It is not possible to prove validity. Instead, validation is "the process by which a test developer or test user collects evidence to support the types of inferences that are to be drawn from test scores" (Crocker and Algina, 1986). Therefore, various data is presented to investigate the interpretation of the scores and the claim that these interpretations are valid considering the conditions under which they are made. This part is based on and extends the preliminary results previously published in Direnga and Kautz (2019).

After reporting on a preliminary study in Chapter 8 that investigates the validity of the CATS as pre-test, the framework and methods used for the revalidation of CATS as post-test are presented in Chapter 9. The description of the collected qualitative and quantitative data is followed by the presentation of the results in Chapter 11. These are structured according to the different aspects of validity introduced in Section 5.1: content and face validity, criterion validity and construct validity, followed by a special investigation of the most extraordinary item in Chapter 12. The part ends with a discussion and conclusion in Chapter 13.

8

PRELIMINARY STUDY - CATS AS PRETEST

The interpretation of pre- and post-test studies is most intuitive and, if suitable instruments are chosen, also most meaningful when the same instrument is used as pre- and post-test. As the requirements for pre-tests are often different from those for post-tests, not every instrument is applicable as both, pre- and post-test to the same course. This challenge has been recognized by Steif and Hansen (2007), stating that

> *"for many subjects in engineering, while there are certainly concepts in previous courses that are relevant, a test that measures conceptual development adequately by the end of the course may not be a valuable measure at the beginning of the course."*

Pre-test instruments should provide a reference by measuring which of the concepts to be learnt are already correctly understood by the students before instruction. Often, familiarity with technical terms and symbols is not required to have a correct conception of the concept to be tested. For example, one might have a solid understanding of the lever principle without knowing the term "couple" or "moment". On many instruments, this familiarity is yet required to correctly interpret and answer the test question. Depending on the research question and the context of the discipline, the (un-)familiarity with technical terms may or may not be targeted by the measurements. In case (un-)familiarity with technical terms and symbols is not part of the investigation, pre-test instruments are required to use language that is interpretable independent of such terms, as correct understanding and interpretation of the test questions is essential for a valid measurement. Unlike on the FCI which uses non-technical language, the responses on instruments using technical language will not reflect the students' beliefs as they cannot understand what the items ask for. This assumption had been supported by statements from the test developer saying that the CATS "offers negligible information as a pretest" (Steif and Hansen, 2007, p. 205). For this reason, the CATS had been declared as unsuitable as pre-test before data collection commenced in the context of the study presented in Part III in an introductory engineering mechanics course at a German university.

This assumption could be empirically confirmed for the investigated population in the context of a Bachelor thesis by Geier (2016). For this purpose, think-aloud interviews were conducted at the beginning of fall 2016 with twelve students at pre-instruction level. The following is a summary of Geier's work with focus on the interpretative problems that were discovered when presenting CATS items to

students at pre-instruction level. As the design of the post-test revalidation was partially influenced by the methods and results, they are laid out in more detail than may be necessary otherwise.

8.1 METHODS

HOW TO ASSESS "INTERPRETABILITY" OF AN ITEM To interpret a CATS item, the students must be able to analyze the setting based on the information contained in both, the text and the graphics. In order to do so, they must correctly interpret the terms and symbols used. Furthermore, they must be able to identify what the question asks for. What is *not* required is the identification and application of the correct concepts to arrive at the correct response, or even any response for that matter. Only if the purpose of the test is to make statements about misconceptions, a response should be required. (This is not the case here, as the CATS shall be used for the purpose to compare different pedagogies in terms of the students' gain in conceptual understanding, leaving the misconception diagnosis out of focus.) Nevertheless, the attempt to find the correct response was used as a vehicle to make possible misinterpretations visible. The interpretation of an item is assessed at three levels:

1. Single elements: interpretation of terms and symbols (e. g. support symbols, arrows, free-body diagram)
2. Connections: relationships between single elements
   a) Content-based (e. g. recognizing that arrows symbolize forces.)
   b) Formal (e. g. recognizing that the instructions at the beginning apply to a certain item.)
3. Aim of question (e. g. recognizing that the question asks for the direction of the force on part X by part Y.)

THINK-ALOUD INTERVIEWS Think-aloud interviews are a commonly used method in PER and EER to reveal students' thoughts (e. g. Engelhardt, 2009). Interviewees are asked to verbally express all their thoughts while being presented certain tasks, in this case selected CATS items. The purpose of the interview was communicated to the interviewee. It was thereby made clear that the research object was the test and not the student, and that possible interpretations of the item were of interest and not their performance in terms of a correct response to the question. Each interview lasted about 30 minutes and was recorded. The video recording only shows the interviewee's hands in order to document sketches and notes created during the interview. The interview recordings were transcribed and analyzed with respect to the above-mentioned levels of understanding. After

the interview, the interviewees were given the option to discuss the correct responses to the items with the interviewer.

THE INTERVIEWEES The students were recruited a few weeks before the beginning of lectures via email, facebook groups, an announcement in the mathematics preparatory course before the start of the first semester, and personal contact. In total, twelve interviews were conducted. All interviewees but one acquired their university entrance qualifications in Germany, either in the same year or one year before. One interviewee had finished their secondary education in 2009 in Syria. None of the interviewees had previous vocational training in STEM-related professions. Their self-reported physics knowledge was largely described as basic. Four interviewees identified as female and eight as male. The variety in study programs was quite high. The decision of admitting first-semester students of all study programs to the study, and not only those who would take the course in which the CATS was administered, was based on the assumption that their previous formal instruction on mechanics from high school is sufficiently similar.

ITEM SELECTION To confirm the assumption that the CATS is not suitable as pre-test, it is sufficient to show that students at pre-instruction level have difficulty interpreting selected test items, especially if the identified reasons for the difficulties also apply to the remaining items. With this approach, the understanding of the items could be probed more deeply in the limited amount of time per interview than if the entire 27 items had to be investigated.

The items were selected with the boundary condition that at least one item of each concept category should be probed. First, items with face elements (e. g. graphics including bearings, cords, arrows etc.) and technical terms (e. g. free-body-diagram, moment, reaction force etc.) that could potentially cause difficulties were identified (see Appendix B). Items that showed redundancy in these aspects were removed from the selection. Along the way, insights from the first interviews resulted in further small modifications of the selection, e. g. to probe whether observed difficulties with certain elements only occurred in combination with specific item features.

Due to time constraints, not every interviewee was shown every selected item. After a few interviews, it seemed as if saturation was reached in terms of the items first probed, which provided the chance to probe other items from the selection. Table 2 shows an overview of the items included in each interview and whether or not they could be successfully interpreted.

Table 2: Evaluation of student interpretation of the items on pre-instruction interviews: correct interpretation or minor misinterpretation (●), inconclusive (◐), major misinterpretation (○). (Items not selected for any interview are not listed.)

Item	Pre-Instruction Student Interview											
	1	2	3	4	5	6	7	8	9	10	11	12
1											●	◐
2	○	○	●	◐	○	○					○	
4	○					●	●					
7	●	◐	●	●	●	◐			●			
8										●	●	
9							●	◐	●			●
10	●	●				●			●	●		
13	●	●	●	○	●	○						
14										●	●	
19			●				●	●				
20				●		◐	○	●				
21										●	●	
23										●		●
25				●		◐	◐	◐				

8.2 RESULTS AND CONCLUSION

Various interpretation problems were found on all four levels. Geier (2016) summarizes these as follows:

(a) While the graphical eltements were often interpreted correctly, the meaning of technical terms was frequently unclear.

(b) Specifically, the technical terms "freischneiden" (separating bodies) and "Freikörperbild" (free-body diagram) are both unknown. In the German version, item 2 uses the former term. In combination with the cords in the given system, this term was often misinterpreted as cutting the cords such that part of the blocks would fall down. In item 9, where only the term "Freikörperbild" is used and no cords are involved, the concept of drawing a free-body diagram of part of the system was more often correctly derived from the response options.

(c) The term "moment" frequently caused problems. Often, it was associated with the more familiar variant of the term "Drehmoment" and tightly linked to motion. In contrast, "force" is a familiar term but also often used synonymously to "motion". Analogously, arrows as graphical representations of forces and moments are often misinterpreted as indicating motion. Generally, the explaining text is often not correctly transferred to the graphical representations.

(d) Applying the instructions "static equilibrium" and "negligible gravity forces" is reported to be one of the main problems. Students either did not understand the terms or did not see when they applied to an item.

(e) Modeling and abstraction is difficult for pre-instruction students. While the given systems were most often well interpretable from the graphical representations, some students were distracted by searching for a possible application of the depicted system. Also, the abstract concept of "arbitrary forces" was very difficult to understand.

Because of the variety of encountered problems with the item interpretation, Geier concludes that the CATS is not suitable as a pre-test in its current form and thereby confirms the initial hypothesis. Based on her findings, Geier presents suggestions for an adapted pre-test version, which would target the observed difficulties, e. g.

- Introducing a "no idea" response option to eliminate the need for guessing.
- Explaining or replacing difficult technical terms, e. g. "all bodies are in static equilibrium" $\rightarrow$ "all bodies are and remain at rest". Although the latter statement is not entirely equivalent to the former (the bodies could also be moving at constant speed), it would help students in interpreting the items.
- Repeating parts of the instructions where they apply.
- Allowing more time for completing the test in pre-instruction administrations.

Even if a CATS pre-test was not readily available for use in the realm of this dissertation it would be valuable for future investigations.

9 METHODS: THE REVALIDATION FRAMEWORK

After having briefly discussed the validity of the CATS as a pre-test in the previous chapter, the following chapters focus on revalidating the CATS as a post-test. The revalidation is largely based on the framework proposed for the validation of CI-claims by Jorion et al. (2015) and the procedure for development of a standardized test described by Adams and Wieman (2011). Jorion et al. (2015) suggest to apply CTT and IRT methods to the entire test to determine problematic items, followed by a structural analysis with problematic items removed. The structure of the instrument shall be first explored with EFA (followed by an optional CFA or DCM). A categorical judgement scheme is provided together with judgement rules to assess the results (see Table 19 in Appendix E). Adams and Wieman (2011) additionally suggest correlations with course outcomes such as exams to add criterion-related evidence. In addition to these rather quantitative approaches, Adams and Wieman (2011) propose to "[c]arry out validation interviews with both novices and subject experts on the test questions" and to "[e]stablish topics that are important to teachers" (Adams and Wieman, 2011, p. 6) for the development of instruments. Although these recommendations are focused on instrument *development*, they are likewise relevant for *revalidation* in a different national context. They touch on the aspects of possible differences in terms of course content and learning goals, as well as language, notations, and conventions. Even though neither of the mentioned frameworks specifically addresses the issue of a different national context, the general goal in terms of validation is the same: ensuring that the content is appropriate, that the items test what they claim to test, and that each item's construct is related to the overall construct being tested. Therefore, most of the suggested methods apply. In addition, a translation analysis will be performed, inspecting the differences introduced by the translation and evaluating those in light of the other differences due to national context.

A detailed description of the methods is provided in the following sections. An overview over the purpose and the implementation of the various analyses is given in Table 3. The presentation of the results in Chapter 11 is structured accordingly.

9.1 INTERVIEW METHODS

As suggested by Adams and Wieman (2011), validation interviews were conducted with both, experts and students. Expert interviews

Table 3: Overview of investigations for establishing validity

Aspect	**Focus question**	**Implementation**
Face validity	Do the items on the test appear to measure the intended construct?	Expert interviews
Content validity	Is the scope of the instrument set on the correct content? Are the concepts tested by the items relevant for local instructors? Do they reflect important learning goals?	Textbook analysis, description of the course, expert interviews
Criterion validity	Is there a correlation with data from an instrument measuring the same or a related construct?	Correlation between CATS and exams
	Are the same distractors effective in both national contexts?	IRC distractor analysis
Construct validity	Does the test measure what it claims to measure?	
	- Can local students interpret the items correctly so that the items can measure the intended construct?	Think-aloud student interviews
	- Did the translation introduce deviations from the original formulations and terms, which changed important aspects of the items?	Translation analysis
	- Are there any problematic items in terms of difficulty and discrimination?	CTT: item statistics
	- Is the scale internally consistent? Do all items add to the reliable measurement of the same overall construct?	CTT: Cronbach's α.
	- Do the items fit the assumption that the probability of a correct response to an item is low for students of low ability and high for those of high ability?	IRT: item characteristic curves
	- Does the test measure precisely across the desired ability range?	IRT: test information curve
	- How many sub-scales does the instrument consist of? Which items group together on the same sub-scale?	Factor Analysis

primarily served to investigate to what extent experts agree that the test addresses appropriate content to measure conceptual understanding of statics. In addition, the interviews provided insight into how experts interpret the items. A correct interpretation of what an item asks for is an essential part of construct validity, and while the interpretation by the *population* (the students) is decisive in this matter, any misinterpretations by experts may also occur among students and must be probed for in the student interviews.

9.1.1 *Expert interviews*

The experts were recruited via first to third-degree personal contact, formal written requests, or at a conference on STEM education. The participants were not offered any incentives. As preparation, the experts were asked to work through the test to become familiar with it. One disadvantage of this approach is that seeing the CATS before the interview possibly introduced a bias before speaking about central concepts (see Section 11.1.1). The major advantage is that it saved valuable interview time for discussion.

The interview protocol contained the questions stated below. These questions were not rigidly addressed in this order. This semi-structured approach allowed the interview to follow the natural flow of the conversation.

1. Do you expect your students to be able to respond correctly to the items?
2. Did you notice any weaknesses in the test or individual items?
3. Which aspects stand out positively to you?
4. Which concepts do you think does the test address? Are they central concepts in the domain of statics?
5. Which central concepts are missing and which ones would you replace or dismiss?
6. How do you assess the choice of distractors? Do you recognize the underlying misconceptions from your teaching experience?
7. How do you assess the level of abstraction?
8. Would you use the test in your course?

The rationale for including each question is as follows:
Question 1: If students are expected to answer the items, the scope of the test does not go beyond the course content, and the items are expected to be interpretable for students.
Question 2 directly asks for flaws which may negatively affect a valid

interpretation of the results. It aims largely at interpretability like question 7.
In general, feedback is often heavy on the negative aspects. Question 3 was included in the protocol to give experts the chance to explicitly mention positive aspects which may otherwise remain unnoticed, and to disclose possible incongruences among the experts' opinions.
Questions 4 and 5 aim at finding out whether the focus of the test is adequate, considering the limited scope. Furthermore, the first part of question 4 aims at verifying that the items actually test what they claim to test ("construct validity").
Question 6: For a valid interpretation of test results, the distractors must be functional, also in the German context. Experienced instructors likely know about typical student errors, some of which are represented in the distractors.

See Section 11.1.3 for results of textbook analysis.

Question 7 was included because of a mismatch in the level of abstraction that was noted in analyzing German and US textbooks. It addresses interpretability of the items.
Question 8 serves as a summative assessment for the test. Depending on the motivation, an expert's intention to use the test can be a sign for quality.

The interviews were audio-recorded, transcribed, and coded with respect to various aspects, such as individual items, the nine CATS concept categories, measurement of expert-like thinking, or expectation of student performance on the CATS. Some of the codes were set a priori, others emerged while reading the transcripts. The range of different responses was very rich, so that not all aspects will be discussed here in detail.

For further analysis, the terms and concepts which were expressed to be central in statics are clustered. For example, one expert may speak of "Newton 3" and another of "actio = reactio", referring to the same concept. Other terms which do not address exactly the same concepts may still be clustered if they are similar enough. For example, "forces on surfaces with negligible friction" may be viewed as a special case of "drawing possible forces". The frequency with which the clusters are mentioned is noted. The more experts agree on the centrality of a concept, the stronger the evidence that this concept is indeed a central one.

9.1.2 *Student interviews*

Through the expert interviews, the issue of face validity, content validity and item interpretability for construct validity is addressed. To investigate whether the items are interpreted correctly by the target population, the students' thought processes had to be made visible. For this purpose, think-aloud interviews were conducted. The post-instruction interview study was designed to resemble the preliminary

study on the CATS's pre-test validity in terms of recruiting interviewees, selecting the items, setting the focus during the interview and the analysis afterwards. This design allows to test the hypothesis that the identified issues in the pre-instruction interviews can be purely attributed to the pre-instruction level of the students. This hypothesis would be supported if the same problems are not observed at post-instruction level.

See Chapter 8 for methods of the preliminary study.

The item selection was based on the selection for the preliminary study, but modifications were made based on results from the preliminary study on the one hand, and the expert interviews, the distractor analysis and the item statistics on the other hand. Items which showed to be already well interpretable at pre-instruction level (i. e. at least three interviews and no problems detected) were removed. The other investigations raised issues also with items which were not already in the item pool. These items were added, resulting in a total number of 15 items to be probed. Details on the motivation for selecting each additional item will be discussed in Section 11.3.3, after the issues found in the other investigations are described.

THE INTERVIEWEES The students were recruited towards the end of the lecture period by an in-class announcement followed by an email which served as a reminder and included more details in form of an FAQ. For comparability, the ideal interview period would have been during the very last week of lectures, where the post-test would usually be administered in class. To make sure that students would be available for interviews, the actual interviews were scheduled a little earlier. This strategy avoided interviews in the busy exam period, even if a second round had been necessary.

As in the preliminary study, each interview lasted about 30 minutes. Again, the purpose of the interview was communicated to the interviewee, and it was thereby made clear that the research object was the test and not the student, and that possible interpretations of the item were of interest and not their performance in terms of a correct response to the question. In the beginning, the interviewee was asked to provide some background data, i. e. their study program, high school-related specifics (GPA, physics level and federal state), any vocational training before their current studies, and their expected mechanics exam grade. The responses were used to assess the heterogeneity of the sample. It was explicitly stated that providing this information is not required to take part in the interview, but none of the interviewees objected.

The interviewees were asked to verbally express all their thoughts while being shown selected CATS items. The CATS instructions (the first page of the test) were shown, followed by two to six individually selected items until the announced time of 30 minutes per interview was up. The video recording only shows the interviewee's hands in

order to document sketches and notes created during the interview, an additional audio track was recorded with an independent device for backup. The focus of the interviews was laid on whether the items are interpreted as intended, which strongly influenced the amount and types of interventions by the interviewer. After the interview, the interviewees were given the option to discuss the correct responses to the items with the interviewer in order to avoid confusion. In some cases, interesting aspects were addressed in this post-interview discussion because the interviewer now asked different questions (often in form of Socratic dialogue) as the focus changed. Transcripts of this are not available because the discussion happened off-record to emphasize that the official interview part was over.

9.2 TRANSLATION ANALYSIS

The original version 4 of the CATS was compared to the respective translated version. One of the following four codes was applied to each deviation by two independent coders[1]:

- missing in the other language
- deviating phrasing
- deviating formatting
- deviating text in the image

Subsequently, these deviations were grouped into categories reflecting possible motivations or justifications, which are discussed in Chapter 11. Personal communication with the leading translator provided additional background information.

9.3 STATISTICAL ANALYSES

9.3.1 *Classical Test Theory methods*

CTT is based on the theory of measurement errors. The central assumption to CTT is that the measured test score X of a person P on an item i is composed of the true score τ and a measurement error ε.

$$X_{P,i} = \tau_{P,i} + \varepsilon_{P,i} \tag{2}$$

The error is assumed to be random and uncorrelated, not only to the true score, but also to the error on a different item answered by the same person as well as to the error on the same item answered by a different person. This implies, for instance, that CTT does not apply to instruments with items which cannot be answered independently,

1 the author and a student assistant

e. g. items where the response to one item is required for answering the other.

Under the assumption of homogeneity of the scale, i. e. that all items measure the same construct, item scores can be added to form a total test score:

$$X_P = \sum_i X_{P,i} = \sum_i \tau_{P,i} + \sum_i \varepsilon_{P,i}. \tag{3}$$

This assumption is also known as the unidimensional τ-*equivalent model*. To meet the assumption, items which do not add valuable information with regard to the construct to be measured should thus be removed. From a purely statistic point of view, increasing the number of items on an instrument increases the expected precision of the measurement because the errors will eventually cancel. Considering that time is a valuable resource and that any test taker sooner or later shows signs of fatigue, tests should only be as long as necessary to measure with the desired precision, which requires items of good discriminatory power and of varying difficulty.

9.3.1.1 *Reliability and Cronbach's α*

Reliability is a central concept of CTT and quantifies an instrument's measurement precision. It is defined in CTT as the ratio of the variance of the true score and the variance of the measurement:

$$\text{Rel} = \frac{\text{Var}(\tau)}{\text{Var}(X)}. \tag{4}$$

It follows from Equation (2) that Rel can only take values between 0 (measures pure error) and 1 (measures pure true score). For the purpose of evaluating instruction, moderate reliability coefficients as low as 0.70 are sufficient (Engelhardt, 2009, p. 24). Although the individual measurements may be subjected to non-significant random errors, these errors diminish in any averaging process when assessing groups.

As the true score is unknown so is its variance, but there are techniques to estimate its size. Most often, a correlation between two tests is applied. There are four main methods:

1. *Parallel test reliability:* If two tests exist which measure the same trait and have the same true scores and error variance, the correlation of both test results administered to the same sample can serve as an estimation of reliability. True parallel tests are rare, especially those measuring personal traits, while parallel forms of performance tests are easier to generate. For instance, calculation ability can be tested with the same items but with different given values.

2. *Retest reliability:* The same sample is tested twice with the same instrument. This approach is often not feasible in practice. Furthermore, the observed trait is assumed to be the same at both measurements, which cannot be guaranteed.

3. *Split-half reliability:* A test instrument consisting of multiple items can be split in half to correlate one half of the items with the other. A requirement is that the halves are similar enough to be seen as parallel tests. As the reliability increases with test length, the splitting must be corrected for by the Spearman-Brown-formula: $\text{Rel}(x) = \frac{2 \cdot \text{Corr}(x_a, x_b)}{1 + \text{Corr}(x_a, x_b)}$, where a and b are the test halves. The split-half procedure only generates a good estimation if the items all measure the same trait.

4. *Internal consistency:* If the items all measure the same trait, the test cannot only be split in half, but in as many parts as there are items. The internal consistency is expressed in terms of Cronbach's α. If k is the number of items on the test, $\text{Var}(x_i)$ is the variance of item i, and $\text{Var}(x)$ is the variance of the entire test, then (see e. g. Engelhardt, 2009)

$$\alpha = \frac{k}{k-1} \cdot \left(1 - \frac{\sum_{i=1}^{k} \text{Var}(x_i)}{\text{Var}(x)}\right). \tag{5}$$

In this investigation, reliability will be assessed using Cronbach's α. It is often incorrectly concluded that a high value of α proves that the scale *is* unidimensional. Instead, the unidimensionality of the scale is an assumption for using Cronbach's α, which only informs about the *extent* of inner consistency of the scale (Moosbrugger, 2012, p. 133). Cho (2016) criticizes the inconsistent naming and misuse of reliability coefficients, most of all Cronbach's α ("Alpha's habitual use is a matter not of mathematics but of marketing"). She proposes "tau-equivalent reliability" as a more meaningful name that emphasizes that Cronbach's α is rooted in the unidimensional tau-equivalent model.

Cronbach's α is sensitive towards the number of items. Adding more items increases α even though individual items may still test an alien construct. As tests should only be as long as necessary, items that do not lead to a definite *gain* in internal consistency as measured by α should be removed.

Instruments are often composed of sub-scales. In this case, the tau-equivalent assumption for the entire test may be violated if not all of the sub-scales contribute to the measurement of one higher-order scale. While Cronbach's α could be very low for the entire test, the sub-scales could exhibit high values of α. This aspect must be considered when evaluating a test designed to measure more than one

scale. In case all of the sub-scales contribute to the measurement of one higher-order scale, the tau-equivalent assumption is not violated.

In general, Cronbach's α (or the tau-equivalent reliability) underestimates the true reliability as defined by Equation 4, if the assumption of a tau-equivalency is violated. Therefore, the value of α can in any case be interpreted as a lower bound of reliability (Moosbrugger, 2012).

9.3.1.2 *Difficulty index*

The difficulty index is a property of an item i. It is given by the fraction of test takers who responded correctly and consequently takes a value between 0 and 1 (see e. g. Engelhardt, 2009):

$$\text{diff}_i = \frac{n_{i,\text{correct}}}{n_{i,\text{correct}} + n_{i,\text{incorrect}}} \tag{6}$$

Contrary to ordinary usage of the word "difficulty", a lower difficulty index shows that fewer students were able to answer the question correctly. Interpretation of the index as an "item score" over all test takers can help to avoid this confusion.

Items which are either never ($\text{diff}_i = 0$) or always ($\text{diff}_i = 1$) answered correctly by the population do not add information and should thus be removed or reworked. Good values for item difficulty are said to be between 0.2 and 0.8 (Jorion et al., 2015).

9.3.1.3 *Discrimination index*

The discrimination index is a property of an item that indicates its power to discriminate between students with high and low total scores. Kelley (1939) proposed to compare the item scores by the 27 % high scorers (H) to the one by the 27 % low scorers (L). The item discrimination can thus be calculated from performance-subgroup difficulty indices as follows

$$\text{disc}_i = \text{diff}_{i,H} - \text{diff}_{i,L} \tag{7}$$

It can take values on the interval from -1 to $+1$, where negative discrimination should be avoided. Ideally, discrimination should be above 0.2 (Jorion et al., 2015).

9.3.1.4 *Shortcomings of Classical Test Theory*

One consequence of the CTT-definition of reliability in terms of variances is that the estimated values depend on the population, regardless which estimation technique is used. According to the population dependence, the reliability over the joint population of all cohorts will be larger than the ones over only the weaker or only the stronger cohorts. This is because the variance of scores is larger for the joint

population. Consequently, the precision of a person's test result must depend on whether the joint population or the cohorts of the same instruction type were chosen for the estimation of reliability, which is illogical. In theory, this should not be the case as the measurement errors are assumed to be random. The concept of reliability in CTT is hence a global one in the sense that there is one reliability for the entire testing procedure including all scores ever generated (Moosbrugger, 2012, p. 138). This concept is obviously flawed as all test scales are limited so that floor or ceiling effects must occur. The reliability in the extreme score ranges cannot be the same as the one in the middle range. Therefore, reliability estimation using CTT can provide sensible results for the middle score ranges, but should be used with care for scores towards the extreme ends (Moosbrugger, 2012, p. 139). Here, IRT models allow for more accurate and differentiated estimations.

Nevertheless, CTT methods provide valuable information about reliability, item difficulty and item discrimination in a fixed test instrument. If items are exchanged, added, or removed from the instrument, it generally has an influence on the other item parameters, which makes CTT less attractive than IRT for the development phase of tests, but for validating complete instruments, it serves the purpose.

9.3.2 *Item Response Theory methods*

IRT is a probabilistic approach to quantify the characteristics and quality of items and tests. Unlike CTT, it does not make an effort to single out a supposedly true score from the erroneous measurement, but instead describes the probability of answering correctly as a function of a person's ability in terms of the measured construct. A test composed of calibrated and well selected items can then be used to estimate the test-takers' abilities. Through calibration, IRT characterizes the item properties like difficulty and discrimination independent of the composition of items on the test and the sample of test-takers (but not independent of the population). This feature makes it very attractive for test developers or adaptive testing, but it can also inform about the qualities of a completed standardized instrument.

9.3.2.1 *Logistic regression models*

The estimated probability $P_i(\theta)$ for a correct response to item i as a function of the test-taker's ability θ (in terms of the construct being measured) is described by the item characteristic curve (ICC). Commonly, they are based on logistic models because they allow to restrict the estimated curves to the interval $[0, 1]$, which makes them suitable models for describing probabilities. Depending on the required com-

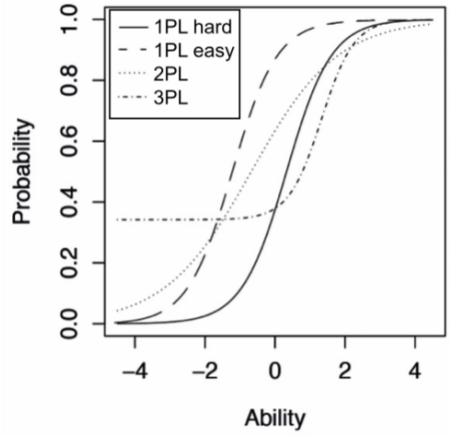


Figure 11: Illustration of the different logistic models in terms of their ICCs. Taking the "1PL hard" curve as a reference, the "1PL easy" curve has a lower difficulty parameter (meaning "easier", unlike CTT difficulty indices) that shows as a shift towards the lower abilities while the shape of the curve is the same. The "2PL" curve has also a lower difficulty parameter but also a lower discrimination parameter, which shows in the comparatively smaller slope at the inflection point. Only the 3PL curve has a non-zero guessing probability, and may additionally vary in the other two parameters.

plexity of the model for adequately describing the data, a one-, two, or three-parameter logistic model may be chosen.

$$P_i(\theta) = c_i + (1 - c_i) \cdot \left(\frac{1}{1 + e^{(-a_i(\theta - b_i))}} \right) \tag{8}$$

Equation (8) describes the ICC of item i using the most complex three-parameter logistic (3PL)-model. The parameters are estimated by fitting the model to the data by maximum likelihood estimation. The *difficulty* parameter b_i is defined as the ability θ at which $P_i = 0.5$. Students with abilities higher that the item difficulty are likely to respond correctly, while those with abilities below the item difficulty are likely to respond incorrectly. The *discrimination* parameter a_i is expressed as the slope at the inflection point of the item characteristic curve. A steep slope indicates a good discrimination between students with abilities below and above the difficulty parameter of the item. The *guessing* parameter c_i is sometimes used to better describe data from closed-ended items, where responding correctly by mere guessing is quite likely. In those cases, the probability asymptote at low ability levels is allowed to differ from zero.

Item characteristics described by the one-parameter logistic (1PL)-model (also called Rasch model) only vary in terms of difficulty ($a_i = 1$, $c_i = 0$), 2PL-models additionally include variance in discrimination (only $c_i = 0$). Figure 11 illustrates the different models.

What becomes evident by the description above is that item characteristics and student abilities are measured on the same scale. When

composing a test instrument from a pool of items, this feature allows to select items with suitable difficulty in order to discriminate most efficiently among those ability levels of interest. For example, a placement test for honor students must consist of different items than a test for identifying students requiring special support.

A useful tool for this matter, but also for validation, are the item information functions $I_i(\theta)$ and the test information function. They can be used to identify at which ability level the item or the entire test provide the most accurate measurement. As such, the item information function is the reciprocal of the measurement variance

$$I_i(\theta) = \frac{1}{\mathrm{Var}(\theta)}. \tag{9}$$

It can be written in terms of the item parameters and Equation (8). For the 1PL (with $a_i = 1$) and 2PL-models

$$I_i(\theta) = a_i P_i(\theta)\left(1 - P_i(\theta)\right), \tag{10}$$

and for the 3PL model

$$I_i(\theta) = \frac{1}{\mathrm{Var}(\theta)} = a_i^2 \left[\frac{1 - P_i(\theta)}{P_i(\theta)}\right]\left[\frac{P_i(\theta) - c_i^2}{1 - c_i^2}\right]. \tag{11}$$

For a single item i, the greatest precision always occurs at the ability level corresponding to the item's difficulty parameter, thus, the item information curve has its maximum at $\theta = b_i$. For the entire test, the test information curve is then given by the superposition of all item information curves:

$$I(\theta) = \sum_i I_i(\theta). \tag{12}$$

A pronounced peak at a certain ability level with otherwise low information values would indicate a suitable test for discriminating between students above or below a certain cutoff ability, such as university entrance examinations. For a valid quantification over a wide range of abilities, such a pronounced peak is undesirable. Instead, a constant high value should be pursued. Further details may be found for example in Baker (2001).

9.3.2.2 *Item response curves for distractor analysis*

The typical multiple-choice/single-response tests are graded dichotomously (correct or incorrect). For CIs, which are also used to learn about student misconceptions, this information is not enough. Additional information on the attractiveness of distractors is required, resulting in non-dichotomous data. Polytomous data from any set of responses that can be ordered on some scale from low to high, such as Likert-scale or partial credit items, can also be modeled by IRT. However, while there are items on the CATS where one distractor may be

"more false" than another one, there is no item for which a complete hierarchy of the severeness of false responses can be established. For example, in item 8, recognizing either one of the two forces as impossible could be rated "less incorrect" than the response that both situations are possible. Ranking the two "less incorrect" options, however, cannot be justified as easily. Therefore, the polytomous IRT analysis is inadequate here.

To obtain information about not only the probability of the correct response but the probability of all possible responses, Morris et al. (2006) suggest a simplified version of ICCs which they call item response curve (IRC). The IRCs describe the probability of a certain response to an item depending on the test score as a proxy for the latent variable. This representation not only provides information on the attractiveness of the distractors for students of different ability, it also allows for blank responses to be considered (which is a problem to be dealt with in this study). Note that with this simplified version of ICCs, the item parameters should only be discussed qualitatively. Furthermore, a guessing parameter as in the 3PL-model cannot be estimated because there cannot be an asymptote on the finite test score scale.

The CATS test score scale theoretically ranges from 0 to 27 points but no students obtained a full score. The resulting test score scale of 0 to 26 is divided equally into 9 bins to smooth out the curves. For each bin, the fraction of students choosing the respective option is plotted. As blank responses are also considered, the probabilities of all options should sum up to 1 for all bins. One example is shown in Figure 12 for item 2. The correct response (d) follows an s-curve which is characteristic for a 2PL-model. The dominant distractor is (c) (IntF). Another example is shown for item 17. This item is very difficult as can be seen from the late rise of the correct response curve (d) (moment and non-normal force both impossible). Distractor (e) (not enough information) is generally ineffective. The other distractors, (including blank responses), are about equally attractive for the very low test scores. With increasing test score, most distractors become less attractive but distractor (c) (non-normal force is possible) becomes more attractive in the middle of the test score range. This pattern indicates that low scoring students tend to guess while higher scoring students understood at least that moments are impossible to transmit in such a setup of a low-friction single-point contact, even if they still believe that the force is possible despite it having a tangential component. Only the very high scoring students also see that the proposed force is impossible if friction is neglected.

See description of distractors in Section 6.1.

In Section 11.2.2, this IRC analysis will be applied to all items and results of this distractor analysis are compared to literature. As the IRC method has not been applied before to CATS data, the comparison to literature must be limited to comparing the overall most attractive

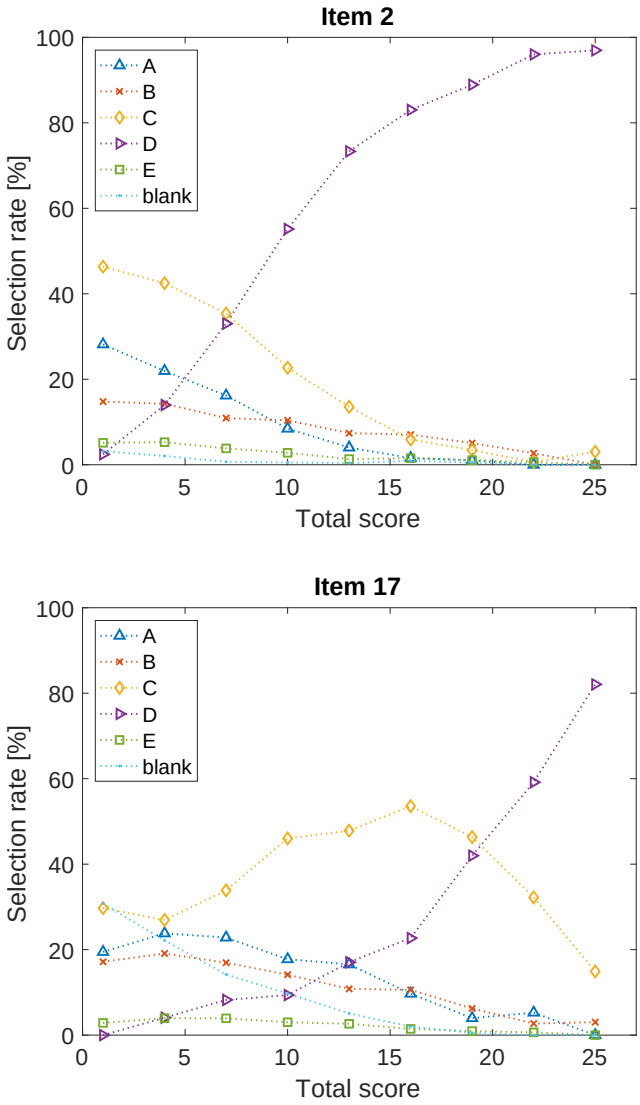


Figure 12: IRC examples

distractors, which have been reported previously. The IRC distractor analysis is categorized as a contribution to criterion validity in the overview in Table 3. While the results of the analysis are not compared to a different instrument, they are related to the administration of the same instrument to a different population. The CATS was developed based on frequently observed errors in the US student population. Similarities in the frequently chosen distractors would indicate that the populations may be similar and therefore the CATS may be equally valid in both contexts.

9.3.3 *Correlation with exam performance*

> *"A concept inventory is of benefit if it gauges levels of understanding that are pertinent to performance elsewhere. Ideally, one would like scores to indicate whether the tester is prepared to apply the concepts during authentic use of the subject in an engineering context. As an admittedly weaker test of this relation, one can seek to compare scores with performance in a Statics course."*
>
> (Steif and Hansen, 2007)

The purpose of the correlation analysis is to show that the CATS does not measure *completely* different variables than the exams, as illustrated by this quote. Correlation analysis is a suitable method for this objective as it is a measure for the relationship between variables. It can be assumed that the Mechanics exams as well as the CATS both measure some sort of knowledge and understanding in Statics. Constructs like intelligence, stamina or motivation are most likely also affecting the measurements of both instruments. While the correlation analysis cannot show *what* the common variables are (this must be hypothesized by theory), positive correlations with other tests measuring a related construct add evidence to the validity of the investigated instrument.

However, there are differences between the CATS and the exams which may reduce the strength of a possible correlation. The CATS measures *conceptual understanding* in the relatively *small realm* of nine concepts. The understanding of each of these concepts is measured repeatedly by three items for better precision. It is a tested, standardized instrument, which is also used for the purpose of detecting common misconceptions. The exam ideally measures how well a student fulfills the learning objectives of the *entire course*. It focuses to a lesser extent on conceptual understanding but largely tests (mostly routine) problem-solving abilities. The exam problems are usually more complex tasks than the qualitative items found on typical CIs, although the complexity is reduced by usually giving the problems already broken down into small steps for easier grading. One major purpose of final exams is to gauge whether a student's knowledge and abilities in the

field are "good enough" to pass. The design of exams hence largely drives student learning. Unlike CIs, exams are not rigorously tested or standardized as they must be different for every administration. Still, they usually are very much comparable from year to year and often based on years of experience in instruction and testing.

Despite the differences between the exams and the CATS, the variables they claim to measure have a common subset. Therefore it is likely that the scores correlate positively, but the correlation is not expected to be very strong. A moderate correlation would suffice to serve as evidence for validity of the CATS.

For the purpose of comparing the results to other literature, Pearson's r is chosen as correlation coefficient. It does not require normally distributed data, but is sensitive towards lack of symmetry in the data distribution, because it is based on covariance:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \tag{13}$$

where σ_{xy} is the covariance of two variables

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{14}$$

and σ_x and σ_y are the respective standard deviations.

To ensure that the distributions of the sampled data are sufficiently symmetric, histograms of the exam and CATS scores will be consulted in Section 11.2.1. Since not all student data could be linked, these histograms will furthermore be used to rule out a significant sampling bias.

9.3.4 *Factor analysis*

Factor analysis is a collective term for a variety of methods with the goal of finding a structure in an underlying set of variables (Bortz and Schuster, 2010). Based on correlation among the variables on this set, the variables are grouped into so-called factors, thereby reducing the complexity of the data. Factor loadings indicate how strongly a variable is correlated with a factor. The analysis *cannot* identify which construct the factor represents, this must be hypothesized by non-statistical investigative methods or theory on the content.

Typical methods for extracting the factors are *principal component analysis (PCA)* and *EFA*. The differences are subtle in that PCA analyzes all types of variance in a variable, while EFA considers covariance only. The former is more adequate for descriptive purposes while the latter should be chosen for inferential statistics (Bortz and Schuster, 2010, p. 427). The concept behind both approaches is that a persons' test result can be represented as a point in a p-dimensional space, where p is the number of variables (in the case of the CATS, $p = 27$). PCA/EFA finds a new space by rotation with respect to the original axes such that each

new axis explains the maximum amount of variance. It can be (mentally) visualized as a sequential process, where the first new axis is fixed in the direction of the maximum variance of the p-dimensional variable-space. This will be the first factor. Variables may be assumed to be correlated or uncorrelated. In case of uncorrelated variables, the second axis must be orthogonal to the first and be fixed in the direction of the maximum variance in the remaining $(p-1)$-dimensional variable-space. This procedure is repeated until the last axis is fixed by the condition to be orthogonal to the others. Here, variance is minimal. In case of correlated variables, the condition of orthogonality does not apply.

Factor analysis does not deliver one well-defined structure as a result. Instead, there are many parameters that can be varied in order to heuristically find the structure with the most explanatory power for the specific context. The number of *relevant* factors is one of those parameters. The number of factors is always the same as the number of variables, but not all factors are equally relevant. The extreme cases are perfect correlation between all variables or no correlation between any of the variables. In case of perfect correlation, there is only one relevant factor that explains all the variance in the data. Translating this situation to a test instrument would mean that any single item would be sufficient to measure the construct. In case of no correlation, there would be as many relevant factors as variables. For a test instrument, that would mean that all items measure entirely independent constructs. Neither case is likely to occur with empirical data, therefore, PCA/EFA will result in data reduction. Context knowledge and theory should inform the decision made by the user how many relevant factors to assume in order to explain most of the variance in the data. One commonly used statistical method that may be applied to help with the decision is *parallel analysis* where the eigenvalues of the empirical correlation matrix are compared to the ones of randomly created data. Factors are assumed to be relevant if the associated eigenvalue is larger than for random data.

The parallel analysis and exploratory factor analysis in this chapter are performed in *R* with the functions *fa.parallel* and *fa()* from the *psych*-package using the oblique rotation method *'direct oblimin'* (Revelle, 2019). Oblique rotation allows the factors to correlate, which is an appropriate assumption here due to the single scale of conceptual understanding of statics and its high internal consistency. Following Jorion et al. (2015), low factor loadings of less than 0.30 are omitted.

See results in Section 11.3.4.2.

10 DESCRIPTION OF THE DATA

In this chapter, the samples of the qualitative and quantitative data that were collected are described, and the test administration procedure of the quantitative data collection is addressed.

10.1 QUALITATIVE DATA

The qualitative data collected as part of this study consists of interviews with experts and students.

10.1.1 *Expert interviews*

Twelve instructors from six different German institutions (Hamburg University of Technology, Frankfurt University of Applied Sciences, TU Darmstadt, Uni Duisburg-Essen, Hamburg University of Applied Sciences, and Technical University of Munich) agreed to participate in semi-structured interviews. The interviews lasted between 45 and 90 minutes and were usually conducted in person in the offices or meeting rooms of the experts; only one interview was conducted in an online meeting. One of the interviews was excluded from the analysis because the interview raised doubt about the expertise of the interviewee (E11). All but one experts are engineers themselves; expert (E2) has a physics background but also multiple years of experience in teaching engineering mechanics.

To correctly interpret the data, it should be mentioned that the expert interviews were difficult to conduct. Many of the experts struggled to (1) think in terms of concepts instead of content, (2) see value in a test that does not cover all the course contents, (3) see the different purpose of the CATS compared to an exam, and (4) refrain from criticizing the items on a very detailed level.

10.1.2 *Student interviews*

In addition to the experts, 16 students at post-instruction level participated in individual think-aloud interviews. The sample was heterogeneous in terms of the recorded characteristics (gender, study program, native language, previous education, and their final high school grade), so that the population is adequately represented. Refer to Table 5 in Section 11.3.3 for the selection of CATS items discussed in the individual interviews.

10.2 QUANTITATIVE DATA

The quantitative data consist of student responses to the CATS and exam data from two cohorts.

10.2.1 *CATS post-test*

The CATS data stem from several cohorts of the largest introductory mechanics course at TUHH. It was collected over several cohorts as a post-instruction test during the last one or two weeks of lecture. The test was administered in class, usually at the end of a lecture as a paper and pen version. The students were provided with a test booklet containing the test questions and a separate answer sheet. An announcement was made at the beginning, explaining the motivation for the data collection, making clear that, while the test results are a valuable feedback, they would not influence the students' grades but rather serve to improve instruction. The earliest cohort was winter semester 2005/2006 and the latest winter semester 2016/2017[1].

After removing 67 data sets with less than nine responses under the assumption that a test completion rate of less than one third reflects unserious test participation, a total number of $N = 4068$ student data sets remained. Over the years, several variables changed, including the semester in which the course was offered, the amount of content and class time, other mechanics related courses as co-requisites, the instructor and the teaching approach. For the purpose of validating the CATS for the general use in Germany, these changes are welcome because the data stem from one institution only, and the changes help to diversify the sample.

Details on the changes can be found in Section 19.1.

At the beginning of the semester, a different test, the FCI was administered in a likewise manner to serve as the pre-test. In Part III, where the effect of the instruction on the test result will be investigated, the analysis will be limited to matched pre- and post-test data. For the validation, all available CATS data (matched and unmatched) are used. If specific methods or research questions require further filtering, the filtering mode of the data is made explicit.

10.2.1.1 *Discussing the time limit on the CATS administrations*

The developers recommend to allow 60 minutes for the CATS but also reported administrations where less time was given, e. g. 50 to 60 minutes (Steif and Hansen, 2006a), 50 minutes (Steif et al., 2005), or "during a 50 minute class period" (Steif and Dantzler, 2005), which probably results in a net time of 45 minutes or less on the test. Steif and Hansen (2007) even report that

1 The author joined the research group in March 2014 and hence collected only the later part of the data.

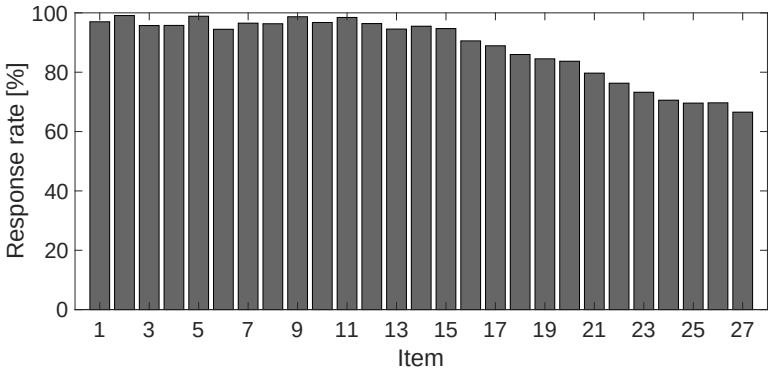


Figure 13: Response rate for each item.

> *"[t]here is no noticeable pattern in the variation of mean scores for the remainder of examinees with time above 25 minutes; [...] Thus, the normally suggested testing time of one hour is plenty of time in which to complete this test."*

The data in this study was collected with a time limit of only 30 minutes. The decision for this time limit was made before the first administration and was then fixed for comparability among the cohorts. The reasons for choosing this time limit are not documented. Possible and not mutually exclusive scenarios could be (1) unawareness of the 60-minute recommendation on part of the test administrator, (2) confusion with the time given for the FCI, and (most likely) (3) a time restriction imposed by the instructor for access to their students.

The strict time limit supposedly had the unwanted effect that students often could not complete the entire test. A plot of the answer distribution reveals up to 5% blanks among all early items, but after item 15 the number of blank responses increases nearly linearly up to about 35% blanks on the last item (see Figure 13).

10.2.1.2 *Discussing the choice of data selection*

As mentioned above, the standard data set on which the revalidation is based contains all available CATS data from the introductory mechanics course at TUHH, regardless of matchings with pre-test data, cohort or type of instruction. The rationale behind grouping all cohorts together is that the CATS is validated mainly with the purpose of discriminating between groups of different instructional methods, as opposed to discriminating among individuals within such groups. If large differences between the test performances of different groups occur, the differences between the groups and the differences between individuals within the groups cannot be equally well measured with the same instrument, just as it is not possible to discriminate equally well between the weights of several heavy object such as cars and between the weights of several light objects such as seatbelt buckles with the same scale.

Another choice to make is how to handle blank responses in the quantitative part of the validation analysis. A blank response is graded as incorrect, even though it is unclear if students even got a chance to view the item. It must hence be assumed that the later items will be overestimated in terms of difficulty. Grading blank responses as incorrect assumes that the item was viewed and the student was not confident enough to give a response. The only alternative explanation for blank responses is that students did not view the item, in which case, in fact, nothing can be concluded about the students' ability to solve it. This scenario is often the more likely one, considering that the distribution of blank responses shown in Figure 13 does not correlate with the difficulties of the items. The choice of grading blank responses as incorrect thus displays the lower performance boundary.

These choices are disputable. To give insight into possible alternatives and to judge the effect of these choices, the revalidation is performed on three alternative data sets, each differing in one aspect from the standard data set:

1. traditional cohorts only ("Trad", N = 1323)
2. interactive cohorts only ("Tut", N = 1496), and
3. only student data with full responses (noBlanks, N = 2010).

The first two alternatives investigate whether the validity depends on the instruction type. These sets do not intersect. The third alternative is a subset of the union of the former two, containing 922 students from the Trad and 1088 from the Tut subset. It allows for all items to be examined under more equal conditions, but the results are biased by a student sample with a certain kind of test-taking behavior (probably fast-working and good at time-management), i. e. the sample is more homogeneous than the entire population.

Most of the quantitative results will be compared to the results presented by Jorion et al. (2015). Their sample of N = 1372 students consists of aggregated CATS data provided by instructors from about 20 institutions. The aggregation results in high heterogeneity with respect to e. g. pedagogy, courses, or time and method of administration. The most appropriate comparison will thus be the one to the standard data set or the noBlanks data set, as Jorion et al. (2015) do not mention how they handled blank responses.

10.2.2 *Exam data*

For details on the cohorts, see Section 19.1.

Exam data from two different cohorts (5 and 7) are available for analysis[2]. The cohort 5 dataset consists of 341 CATS and 420 final exam scores, of which 305 could be matched. The cohort 7 dataset consists

2 Both cohorts used Tutorials, hence no distinction with respect to teaching approach can be made.

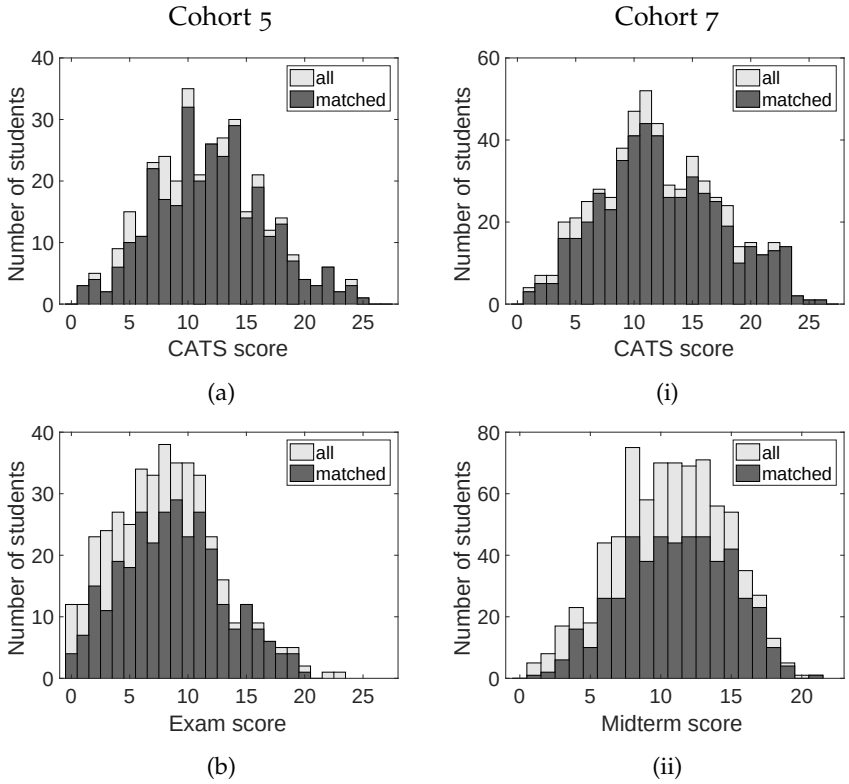


Figure 14: Histograms of CATS and exam scores for two cohorts (cohort 5 in the left and cohort 7 in the right column). The sample for which CATS and exam scores could be linked (matched) is a good representation of the population (all).

of 566 CATS and 766 midterm exam scores, of which 497 could be matched. The maximum possible scores were 27 points on the exam in cohort 5 and 22 points on the midterm in cohort 7.

Histograms of CATS scores and exam scores are shown in Figure 14. Separately indicated are the proportions of students for whom the two tests, CATS and exam, could be matched. These subsamples seem to be an adequate representation of the overall population and their distributions are sufficiently symmetric for using Pearson's r.

11 RESULTS

In the following sections, the pieces of evidence gathered in the validity analysis will be presented. First, evidence gathered from expert interviews, course descriptions and textbooks addressing content and face validity will be shown. Second, evidence from exam correlation and distractor analysis addressing criterion validity are shown. Finally, evidence on construct validity is presented based on expert and student interviews, the translation analysis and the statistical analyses.

11.1 CONTENT AND FACE VALIDITY

To assess whether the CATS tests concepts which are relevant for instructors in terms of conceptual understanding of statics, a content analysis was conducted based on eleven expert interviews with instructors from six different German institutions, course description and textbooks. Face validity was implicitly established in the expert interviews.

11.1.1 *Expert interviews (content)*

The results from the expert interviews will be presented according to the questions on the interview protocol. Only those statements addressing content issues are mentioned here. Statements considering item interpretability are presented in Section 11.3.1.

Do you expect your students to be able to respond correctly to the items?

Most of the experts would expect their students to be able to respond correctly to the items after instruction, as the following two quotes illustrate:

> *"I'm relatively certain of that, yes. And since they are trained, they are faster than we are, hopefully. Well, a few at least. Nah, seriously now, they should be able to [solve this], yes."*
>
> *"Da bin ich mir relativ sicher, ja. Und da die trainiert sind, schneller als wir, hoffentlich. Naja ein paar zumindest. Nee, also Spaß beiseite, müssten die können, ja."*
>
> *E6 #00:57:25-5#*

"[...] if they are not able to answer these questions at the end of the semester, then they should not even be able to take the exam, because it is so fundamental, [...]."

"[...] wenn die das am Ende des Semesters diese Fragen nicht können, dann sollen sie bitte auch gar nicht zur Klausur können, weil das so basic ist, [...]."

E8 #00:52:54-2#

One expert even stated that he would expect his students to be able to answer the questions already after four weeks into the curriculum.These expectations support the claim that the CATS is limited to concepts which are not only part of the curriculum, but also very fundamental and essential.

Some experts voiced the same expectation but expressed doubts that their students would indeed perform well. They seem to have experienced a gap between actual student understanding and course objectives.

"Yes, of course, yes. They should actually be able to answer them [the questions], yes. In principle, yes. Whether they can really answer it, that's another question."

"Ja selbstverständlich, ja. Die [Fragen] sollten sie eigentlich beantworten können, ja. Vom Prinzip her, ja. Ob sie es immer beantworten können, das ist eine andere Frage."

E5 #00:03:06-8#

"I would actually expect the students to be able to answer the questions, at least by a process of elimination. However, I have my doubts as to whether they will actually be able to do that."

"Ich würde eigentlich erwarten, dass die Studierenden die Fragen beantworten, zumindest nach dem Ausschlussprinzip beantworten können. Ich habe aber Zweifel, ob die das wirklich, teilweise hinkriegen."

E3 #00:05:26-3#

One expert disagreed, stating that

"[...] the test requires too much intuition, for my taste."

"[...] mir ist bei diesem Test zu viel Intuition gefragt."

E4 #00:12:16-4#

When asked what he meant by "intuition", he elaborated:

> *"Intuition comes when I have solved 20 of such problems or after I have served as a teaching assistant on this topic and have been asked questions which surprised me, and I had to think about it myself and question my own... yes, constructed knowledge and connections, then it becomes something like intuition. But I think, expecting this from a first-year student is difficult.*
>
> *"Intuition kommt, wenn ich 20 solcher Aufgaben gelöst habe oder schon eine Übung gleitet habe zu dem Thema und Fragen kommen die mich überrascht haben und ich selber mal nachdenken muss und mein... ja, mir gebildetes Wissen und die Zusammenhänge selber in Frage stellen muss, dann wird irgendwas wie Intuition draus. Aber ich glaube, das von einem Erstsemester zu verlangen, finde ich schwierig.*
>
> *E4 #00:14:02-6#*

What E4 describes as "intuition" resembles a description of deep or expert-like understanding, which is indeed what the CATS intends to measure. E4's statement can hence be interpreted as supporting evidence that the CATS measures conceptual understanding.

This link between conceptual understanding and intuition was also mentioned by Montfort et al. (2009) (→ p. 21).

The concern brought forward by E4 seems to be motivated by the notion how an engineer approaches problems. This and other expert interviews emphasized that being able to solve any complex statics problem *systematically* is a central learning goal in engineering mechanics...

> *"[...] if one consistently follows the rules, which most [students] don't do, then one can find out on all the questions what is correct and what is incorrect." [...] "That's what we want to encourage them to do, to not immediately go for the intuitive answer."*
>
> *"[...] wenn man sich da eben ganz konsequent an die Regeln hält, was eben die meisten [Studierenden] nicht tun, dann kann man bei allen Aufgaben rauskriegen was richtig ist und was falsch." [...] "Dazu wollen wir sie ja auch animieren, dass sie nicht gleich gucken, was muss denn rauskommen gefühlt."*
>
> *E3 #00:30:43-5# ff.*

... while the systematic approach should not replace a final check for plausibility of the obtained result.

> *"It is also nice when one has a feeling for the mechanics. Of course I encourage them to think again at the end if what they calculated is plausible."*

> *"Das ist auch schön wenn man ein Gefühl für die Mechanik hat. Ich animiere sie natürlich auch am Ende nochmal darüber nachzudenken, ob es denn sein kann, was ihr da ausgerechnet habt."*
>
> *E3 #00:31:27-1#*

The phrase "feeling for the mechanics" also indicates a kind of "intuition" that seems to be a desirable goal for a professional engineer.

> *"[...] if [later the students] work in a company and they do any calculations, they always have to do plausibility checks anyway."*
>
> *"[...] wenn [die Studierenden später] in einer Firma sind, müssen sie sowieso, wenn sie da irgendwie rechnen müssen, sowieso Plausibilität immer machen."*
>
> *E12 #00:39:15-9#*

E12 then continued that it would be nice to test students on their intuition and then came up with the idea of using non-intuitive examples in his teaching to warn students of relying entirely on their intuition. Summarizing these statements leads to the conclusion that students should learn to ignore their intuition at first and approach any problem systematically. Once a solution is obtained, they are expected to check for plausibility, for example by consulting their intuition.

In some experts' opinion, the CATS and the conditions of test administration do not provide the appropriate frame for applying the systematic approach taught in the course.

> *"So I have one minute per question. That means I cannot draw a free body diagram, I cannot set up equilibrium conditions. This means that it is not required to solve the tasks systematically. But that's exactly what we're teaching."*
>
> *"Also ich habe pro Frage eine Minute. Das heißt ich kann kein Freikörperbild zeichnen, ich kann keine Gleichgewichtsbedingungen aufstellen. Das heißt es wird nicht verlangt, die Aufgaben systematisch zu lösen. Das bringen wir aber gerade bei."*
>
> *E4 #00:12:36-5#*

> *"As I understood it, one answers a lot of questions in a short time and that means that one should actually be able to solve it by only looking at it not by isolating the system."*
>
> *"So wie ich das verstanden hab, macht man ja in einer kurzen Zeit sehr viele Aufgaben und das heißt, man müsste das eigentlich sehen und nicht freischneiden."*
>
> *E1 #00:01:45-7#*

If given the appropriate amount of time, however, E1 would expect his students to solve the questions using the systematic quantitative approach learnt in class. E4 furthermore agrees that "the test is good" and that it addresses known student difficulties, but that it is administered too early in the study program (E4 #00:28:13-2#).

In general, the time given to the students to take the test was frequently criticized as too short. Experts reported that they could not finish the test in the given time frame.

> *" [...] it took me 45 minutes and so in 30 minutes... okay now I am not the fastest but I find that ambitious."*
>
> *"[...] ich habe eine Dreiviertelstunde gebraucht und in also 30 Minuten... okay jetzt bin ich nicht der schnellste, aber das finde ich sportlich."*
>
> *E8 #00:25:11-9#*

Did you notice any weaknesses in the test or individual items?

Many of the answers to this question relate to the issue of item interpretability and are hence discussed in Section 11.3.1. Content-wise, the strong focus of the CATS on roller joints (and likewise pin-in-slot joints) was criticized. E1 does not see the roller as a standard content like trusses or rods. He argues:

> *"We have no chapter on rollers, for example, [...] and I am critical of making it such a central [concept], because, as soon as we have dynamics, it [the radial direction of possible forces] no longer applies."*
>
> *"Wir haben kein Kapitel über Rollen zum Beispiel, [...] und das sehe ich auch kritisch das als so einen Mittelpunkt zu stellen, weil, sobald wir Dynamik haben, das halt nicht mehr gilt[, dass die Kräfte nur radial wirken]."*
>
> *E1 #00:01:20-0#*

According to E1, the students are not supposed to memorize how a roller element behaves in terms of forces, but they should always use the systematic approach, starting with drawing free-body diagrams (E1 #00:01:30-6# ff.). In addition to possible problems in the subsequent dynamics course, cultural motivations for criticizing the concept of rollers were expressed by E2:

see also Section 11.1.3

> *"German mechanical engineering draws a distinction between engineering mechanics and construction elements. The [US-]Americans usually don't do that. [...] Someone coming from [German mechanical engineering] may say: 'A roller joint is a construction element, that has nothing to do with engineering mechanics, so I don't address that here.' [...] But I don't mind it here, since in my opinion, it is only included here because, with the roller, you have a range of possibilities to address topics like isolating systems, equivalence, forces, equilibrium etc. So for me the roller is not the central aspect. It it merely a tool to introduce the mechanics [concepts]."*
>
> *"[...] der deutsche Maschinenbau der teilt ja auf zwischen der technischen Mechanik und den Konstruktionselementen. Das tun die [US-]Amerikaner aber normalerweise nicht. [...] [W]enn man da [im deutschen Maschinenbau] groß geworden ist, dann sagt man: 'Eine Verbindung durch Rollen oder eine Rolle ist ein Konstruktionselement, das hat eigentlich in der technischen Mechanik nichts zu suchen, deshalb mache ich das da nicht'. [...] Ich finds hier aber nicht so schlimm, weil das hier für mich nur drin ist weil man eben mit der Rolle noch mehr Möglichkeiten hat, dass die mal Freischneiden, Äquivalenz, Kräfte, Gleichgewichte etc. zu behandeln. Also für mich steht hier nicht die Rolle im Vordergrund. Also die Rolle ist hier nur [...] Mittel zum Zweck um die technische Mechanik einzuführen."*
>
> *E2 #00:08:54-4#*

E2 disagrees with the opinion, which he often witnessed among his colleagues, that engineering mechanics must be strictly separated from construction elements. One may argue that his view is not representative here as it is the one of a physicist, but it is indeed shared by E10, an expert with a background in mechanical engineering.

> *"It [the concept of the frictionless roller] is of course important. I mean, we need to speak about friction and we simply need to know when there is friction and when there is none."*
>
> *"[...] Es [das Konzept der reibungsfreien Rolle] ist natürlich wichtig. Ich meine, wir müssen über Reibung sprechen und wir müssen einfach wissen wann es Reibung gibt und wann nicht."*
>
> *E10 #00:22:46-7#*

Thus, for both E2 and E10, the objective of the roller joint items is not to test understanding of rollers, but to test understanding of more general concepts.

Which aspects stand out positively to you?

It is human nature to first notice negative aspects, therefore comments on the positive aspects were explicitly prompted.

> *[The students] are forced to vividly recall in their minds what they should have learnt. And this is precisely what the intention is, these false ideas, some of which have been stored in the mind.*
>
> *"[Die Studierenden] werden gezwungen, sich intensiv nochmal das hervorzurufen, im Kopf, was sie eigentlich gelernt haben sollten. Gerade was ja die Intention auch ist, diese falschen Ideen, die man teilweise im Kopf abgespeichert hat."*
>
> *E3 #00:30:43-5#*

This quote supports the hypothesis that the CATS successfully addresses misconceptions ("false ideas") and tests the right concepts ("what they should have learnt") from an expert's point of view. Drawing correct free-body diagrams is regarded as essential among those concepts, and the CATS is perceived as testing this skill, and not only with the concept category that asks specifically for drawing forces on separated bodies:

> *So, basically, if we are talking about drawing free body diagrams - which we attach GREAT importance to - these tasks make a lot of sense.*
>
> *"Also grundsätzlich, wenn wir hierbei sind Freikörperbilder zu zeichnen - wo wir einen GROSSEN Wert darauf legen, [...] - machen diese Aufgaben viel Sinn."*
>
> *E6 #00:00:57-3#*

Furthermore, it was appreciated that the CATS does not follow a "plug-and-chug"-scheme, which relates to the conceptual nature of the items.

> *"[T]hese are nice questions because you have to think a little bit about what to do."*
>
> *"[D]as sind schöne Aufgaben, weil man auch ein bisschen drüber nachdenken muss, was zu tun ist."*
>
> *E6 #00:59:48-5#*

Such a statement also implies that most tasks given to the students on other occasions are indeed "plug-and-chug".

Which concepts do you think does the test address? Are they central concepts in the domain of statics?

To account for the occasional difference in use of terminology among the experts' answers to this question, the mentioned terms and concepts were grouped into concept clusters for the analysis. The following concepts were explicitly named as central and - at least in parts

- to be addressed by the CATS. The numbers in parentheses indicate the number of experts who explicitly named this concept. Concepts which were named only once are not considered. If a concept was not explicitly named as central by an expert, no conclusion is drawn about the expert's opinion on the concept's centrality.

- *Free-body diagrams* (10): defining the system boundaries, drawing the FBDs of the separated parts, replacing removed parts by the respective forces.
- *Equilibrium* (8): stating the conditions for equilibrium, determining reaction forces required for equilibrium.
- *Possible (reaction) forces* (6): determining possible forces, forces on surfaces with negligible friction, making reaction forces visible.
- *Friction* (4): limit of friction, discriminating between kinetic and static friction.
- *Newton's Third Law* (3): action/reaction-principle
- *Couple as a free vector* (3)
- *Static equivalence* (2): replacing one set of loads by another while maintaining equilibrium.
- *Decomposition of force vectors* (2)

As the concepts are not independent of one another, these clusters inevitably show a certain overlap. For example "making reaction forces visible" is also linked to the *Free-body diagram* concept. Likewise, "determining reaction forces for equilibrium" is also linked to the concept *Possible (reaction) forces*.

The strong prevalence of the *Free-body diagrams* concept cluster, followed by *Equilibrium* and *Possible (reaction) forces* indicates that these concepts are the core of statics for German instructors, and they acknowledge the subtle difficulties in these concepts.

> *"[...] And the [CATS] does the job, I think. It comes really close to what we are doing in Mechanics 1, in statics of rigid bodies. Asking for free-body diagrams, drawing possible forces and moments, identifying boundary conditions. . . . Yes."*
>
> *"[...] Und der [CATS] tut den Job, finde ich. Also, der ist eigentlich schon sehr nah dran an dem was wir in der Mechanik 1, also in der Statik starrer Körper, auch tun. Immer mit diesen Fragestellungen: Freischneiden, Freikörperbilder bilden, die richtigen möglichen Kräfte und Momente eintragen, Randbedingungen identifizieren. . . . Ja."*
>
> *E2 #00:06:06-5#*

"In my opinion if they can do what's being tested here and then remember only half of that, that's a lot. [...] isolating systems and drawing free body diagrams, that's simply essential [...] and it's not trivial either. [...] If there is a bit of friction or something is not that important."

"Also ich finde wenn sie das können was hier abgeprüft wird dann und nur die Hälfte von dem behalten, dann ist das viel. [...] dieses Freischneiden und das Bilden von Freikörperbildern, das ist einfach so zentral [...] und ist auch nicht trivial. [...]"

E2 #00:07:09-0#

The proficient use of FBDs is even seen as one of the key characteristics that differentiates an engineer from a physicist.

I: "Can you [...] say what the most important concepts would be, in your opinion?"

E5: "[...] Equilibrium, yes. Free body diagram, that's something essential that distinguishes engineers from physicists. Physicists also know Newton's axioms, but they can't apply them to specific cases. They usually fail because they don't realize where the system boundary actually is [...]. [...] and you have to get that straight by using a free body diagram. And if you can't do that, the equilibrium conditions are useless."

I: "Können Sie [...] sagen, was die wichtigsten Konzepte für Sie wären?"

E5: "[...] Gleichgewicht, ja. Freikörperbild, das ist ganz was essentielles, was Ingenieure von Physikern unterscheidet. Die kennen zwar auch Newtonsche Axiome aber sie können sie schlecht anwenden, ja, auf konkrete Fälle. Und das scheitert meist daran, dass sie sich nicht klar machen, was eigentlich [...] die Systemgrenze ist. [...] und das muss man sich klar machen, anhand eines Freikörperbilds. Und wenn man das nicht kann, dann nutzen einem die Gleichgewichtsbedingungen wenig."

E5 #00:07:34-3#f.

Interestingly, *Newton's Third Law* was only explicitly named by three experts, despite it being one of the four basic axioms of statics. It is possible that this principle is implicitly seen as included in the *free-body diagrams* concept by some experts, as it has to be applied once multiple bodies are examined.

Friction was often assessed as a difficult concept which is also addressed in the course, but the opinions diverge whether it is an essential one.

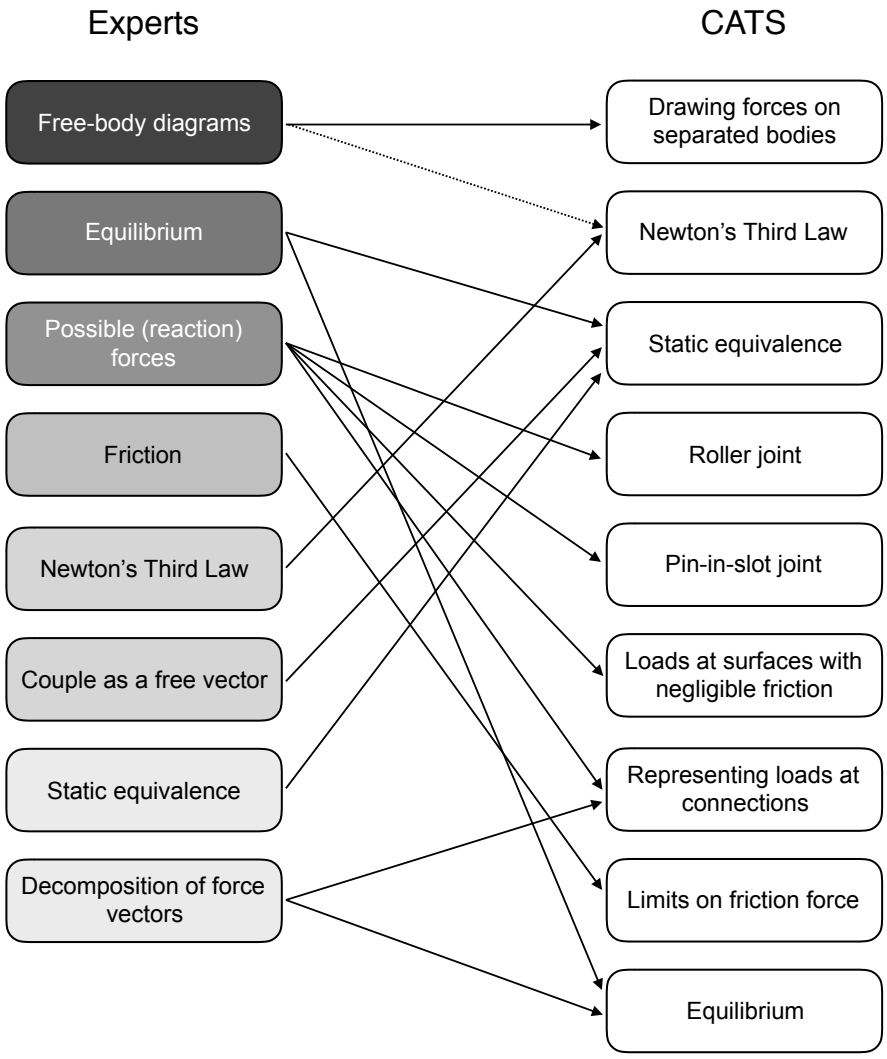


Figure 15: Mapping the concepts, which experts explicitly named as central and to be addressed by the CATS, to the concept categories given by the CATS developers. Darker shading indicates higher frequency of experts naming the concept as central

What follows is an analysis of how many of the CATS concept categories are reflected by the concepts named most often as central by the experts. Figure 15 shows a mapping of these expert-identified concept clusters to the concepts defined by the developers of the CATS. It is based on the author's interpretation of the data. The arrows indicate in which CATS concept categories the concepts named by the experts are addressed. (The dotted connection is only weakly supported by the data.) The mapping is neither injective nor surjective as some concepts named by the experts map to more than one CATS concept and some CATS concepts are mapped to more than one expert concept cluster. The concepts in the top three concept clusters were named by at least half of the experts to be central. (Remember that not naming a concept does not mean that it is not considered to be central.) These three together already map to seven out of nine CATS concepts (one more if the dotted connection is considered). If the fourth, i. e. *Friction*, is included, which was named by about one third of the experts, the number of covered CATS concepts increases to eight out of nine (one more if dotted connection is considered). If the top five are considered, which were each still explicitly named by at least one fourth of the experts, all nine CATS concepts are covered. The centrality of the concepts chosen for the CATS seems to be supported by the experts. On a closer look, however, not all concepts clustered in *Free-body diagrams* are addressed by the CATS. For example, defining appropriate system boundaries for the free-body diagram such that a system of equations can be derived and solved to determine the desired quantity is not addressed by the CATS items. The free-body diagram items on the CATS only ask for which forces should be drawn in a free-body diagram of a given system of bodies. Apart from this, there are further concepts named as central by the experts but not seen as to be addressed by the CATS.

See the four necessary steps to construct FBDs on p. 16. The CATS only tests step 3.

Which central concepts are missing and which ones would you replace or dismiss?

The most repeatedly mentioned concepts to be missing are summarized as follows. Again, the numbers in parentheses indicate the number of experts explicitly naming this concept to be missing on the CATS.

- *Static determinacy* (6): analysis of proper constraining, instability.
- *Vectors* (4): performing mathematical operations on force and moment vectors (e. g. parallelogram of forces, trigonometric functions etc.), understanding moment as a vector product (requires 3D problems).
- *Center of gravity and centroid* (3)

A predictive power of the CATS on these concepts has previously been reported (→ p. 62)

- *Internal forces and moments* (3)
- *Structural analysis* (3): trusses, both two- and three-dimensional, determining zero-force members
- *Virtual work* (2)
- *Determining reaction forces* (2): encompasses the entire process of drawing a free-body diagram of the whole system and calculating values for the reaction forces.

Static determinacy is missed most frequently, followed by vectors. While parts of the *Vectors* cluster mentioned here are acknowledged by the experts to be addressed by the CATS in *Couple as a free vector* and *Decomposition of force vectors*, there seems to be an agreement that vectors are such a central concept for statics, that they should be addressed seperately. This is rather surprising as one may argue that vectors are a *prerequisite* for the Statics course and thus merely a tool that is used to describe the essentials of statics.

A large part of the experts' criticism was directed towards the testing *methods* rather than the *concepts*. Many experts did not clearly distinguish between assessing which concepts are addressed in the CATS and *how* they are addressed. As a result, the focus of the respective interviews shifted from the concept itself to the method by which student understanding of a concept is assessed by the CATS. Consequences of this can be seen in the cluster *Determining reaction forces*. For example, it was often stated that students should be tested on drawing free-body diagrams in an open-ended question format instead of identifying the correct one among multiple-choice options.

> *"So the question [item 2] itself is okay. I would have asked the student: 'Sketch the free body diagram for the system so and so'."*
>
> *"Also die Aufgabe [2] an sich, ja, ist ok. Ich hätte vom Studenten verlangt: 'Skizzieren Sie mir das Freikörperbild für das System sowieso'."*
>
> *E5 #00:08:55-9#*

Likewise, it was stated that students should formally state, i. e. write down, the conditions for equilibrium and show that they can determine, i. e. calculate, correct values for reaction forces instead of analyzing qualitatively whether a body subjected to external loads can or cannot be in equilibrium.

I: "[...] You said here [...] 'Setting up equilibrium conditions and checking the results for problems concerning equilibrium'. Would you, after we have looked at the items 20 and 26 - those were the ones which were both about the equilibrium conditions - would you still say that this [topic] is missing or would this thereby be represented in the test for you?"

E8: "No, [...] [W]hat I mean by this is that one has to write down these equilibrium conditions for example with the correct lever arms, you know? [...] I mean, setting up an equation requires additional knowledge, in my opinion. Or decomposing forces, for example, [...] you have to process that correctly in a calculation, for example, you know? [A]nd I think that this is not yet, not as thoroughly included in the test as one could check it with another question for which one has to write down specific equations."

I: "[...] Sie hatten hier gesagt [...] 'Aufstellen von Gleichgewichtsbedingungen und Kontrolle des Ergebnisses für Gleichgewichtsaufgaben'. Würden Sie, nachdem wir Aufgabe 20 und 26 uns angeguckt hatten - das waren die, die beide nach den Gleichgewichtsbedingungen fragen - würden Sie denn immer noch sagen, dass das fehlt oder wäre das damit für Sie im Test mit drin?"

E8: "Ne, [...] [D]as, was ich darunter verstehe ist, dass man diese Gleichgewichtsbedingungen schon aus hinschreibt mit den richtigen Hebelarmen zum Beispiel ne? [...] also Aufstellen erfordert finde ich noch zusätzliches Wissen, zum Beispiel. Oder Kräfte zerlegen zum Beispiel, [...] das muss man schon auch in einer Rechnung dann richtig verarbeiten, zum Beispiel ne und dann finde ich schon, dass das hier noch nicht so drin ist wie wir das, wie man das abprüfen könnte in einer anderen Aufgabe wo man konkrete Gleichungen hinschreiben muss."

E8 #00:28:25-1# f.

This latter mindset shows that the focus of some experts was less on conceptual understanding than on solving standard quantitative and exam-like problems. The conceptual focus of the CATS does not follow the testing tradition in engineering education. It is therefore understandable if this new approach is criticized.

This mindset was likewise observed in the US, (→ p. 62).

How do you assess the choice of distractors? Do you recognize the underlying misconceptions from your teaching experience?

The distractors were mostly assessed positively in the sense that they address a beginner's misconceptions:

> *"The [distractors] are good. In my opinion, they're well thought out and they're accurate representations of how the students may think incorrectly."*
>
> *"Die [Distraktoren] sind gut. Ich meine die sind durchdacht und die sind tatsächlich meine Meinung, wie die Studis falsch denken können."*
>
> *E10 #01:11:14-1#*

Only some experts did not immediately see the reasoning behind some of the distractors. Even after discussing the distractors of item 7, E12 still struggled with the underlying misconceptions.

> *[regarding item 7]: "Yes, well, maybe [(b)] is even less wrong than [(a)]." [...] "Yes, because [in (a)] I move a moment now, I move a moment, there is now... with a force there is at least a lever arm involved. But that's... I double the moment and move it somehow, that's simply... I find it hard to understand what he was thinking."*
>
> *[bezieht sich auf Aufgabe 7]: "Ja gut, vielleicht ist [(b)] sogar weniger falsch als [(a)]." [...] "Ja, weil ich verschieb jetzt ein Moment [in (a)], ich verschieb ein Moment, da ist jetzt...bei einer Kraft ist wenigstens irgendwo noch ein Hebelarm drin. Aber das ist schon, ich verdoppel einfach das Moment und verschieb's irgendwie, das ist schon einfach... da tue ich mir schwer jetzt noch nachzuvollziehen, was hat er sich denn da dabei gedacht."*
>
> *E12 #00:09:22-8# ff.*

The response distributions are shown in Figure 52.

To be fair, the student data shows that response option 7(a) is not very frequently chosen, unlike 7(d). For E12, response 7(d) is merely a calculation error, which is not as serious as conceptual errors (E12 #00:07:56-1# ff.). The possibility that a student, who selected 7(d), incorrectly ruled out the correct response based on the misconception that the couple must be centered about the point where the 200 Nm couple to be replaced is drawn, does not occur to him. To his defense, it needs to be said that even Steif and Dantzler (2005) wrote that the fact that the equivalent force couple in the correct response of item 7 is not centered at the same point as the couple to be replaced makes the item only "slightly more difficult". The results from Brose and Kautz (2011), however, show contradicting evidence, indicating that students struggle immensely with the concept of a couple being independent of the point of application.[1]

1 They hypothesize that this struggle is likely fostered by predominantly oversimplified symmetric representations in textbooks.

Would you use the test in your course?

Three of the interviewed experts (E1, E2, E4) had already provided course time to use the test in their courses as they were part of the data collection described in Section 10.2, but only E2 was committed to it. Some other experts (E5, E7) did not see additional value in using the CATS in their course. Three others (E8, E10, E12) were interested in using the CATS, mostly with the purpose of a learning instrument, in order to help students learn from the items. One expert (E3) was interested to use the CATS for assessment of instruction.

Summary: Experts on content

Regarding "content validity", the expert interviews revealed that the concepts addressed by the CATS are part of the curriculum. Every concept on the CATS was considered central (or to be contributing to larger central concepts) by at least three experts. The concept of free-body diagrams was especially emphasized. One exception is the strong focus on the roller as a concept of its own. In the German context, the roller joint concept changes its purpose. While it is actually the specific behavior of the roller joint, which is of interest in the US context, the roller becomes a tool to test more general concepts; it acts as an example system containing a frictionless joint. (The same also applies to the "Pin-in-slot" concept.) Regardless of the centrality of this concept, most experts agree that students should be able to answer these and all other items on the test. The distractors align to a large part with the instructors' experience of typical student errors.

The concept most often named to be central but missing was *static determinacy*, followed by *vectors*, a concept which is often a prerequisite for the course. Other concepts perceived by the experts to be central and missing are *center of gravity and centroid*, *internal forces and moments*, *structural analysis*, *virtual work*, and *determining reaction forces*. As the CATS has a limited scope and the addressed concepts were likewise assessed to be central, adding these concepts to the CATS would result in other cutbacks, such as less precise measurements or more time on the test. Such a modification would hence not necessarily lead to a higher quality instrument. One option for a future version of a German CATS would be to exchange the roller concept for items on static determinacy.

11.1.2 *Course description*

The course content of the introductory engineering mechanics course at TUHH is described in the module description (Hamburg University of Technology, 2019) as:

- force systems and equilibrium,

- supports,
- trusses,
- weight force and center of mass,
- friction,
- ropes and chains,
- and internal forces and moments in a beam.

The CATS thus does not contain any content which is not addressed by the investigated course, but it does not necessarily contain the most important concepts. For example, the correct/systematic labelling of forces on a free-body diagram was seen as less important by some experts than the prior step: defining where to separate the bodies. Also, while different types of supports are introduced, the focus is different between German and American engineering mechanics instruction in terms of the degree of abstraction in the depicted mechanical systems, as is laid out in the following section.

11.1.3 *Textbook analysis*

Considering *content* of a typical statics textbook, US and German books are quite similar. The differences rather emerge in the "how" than in the "what". While the number and sequence may vary, German textbooks often introduce the *axioms of statics* explicitly (e. g. Brommundt et al., 2007; Dankert and Dankert, 2013; Gross et al., 2011). US textbooks most often do not refer to these basic principles as axioms, but the concepts are introduced. The importance of free-body diagrams, which also emerged from the expert interviews, is emphasized in both German and US textbooks:

> *"Of special importance here are the principles of separating bodies, action/reaction, and the "free body diagram". They are applied in solving nearly all statics-related problems."*
>
> (Gross et al., 2011, p. 6)[2]

> *"[...], a thorough understanding of how to draw a free-body diagram is of primary importance for solving problems in mechanics."*
>
> (Hibbeler, 2004, p. 195)

2 German original: "Besondere Bedeutung haben dabei das Schnittprinzip, das Wechselwirkungsgesetz sowie das "Freikörperbild". Sie werden bei der Lösung von nahezu allen Problemen der Statik angewendet."

STATIC DETERMINACY Inspecting the central but missing content on the CATS extracted from the expert interviews (static determinacy, vectors, center of gravity and centroid, internal forces and moments, structural analysis, virtual work, and determining reaction forces, → p. 113), the only difference is found with respect to the centrality of static determinacy. Static determinacy addresses the question whether a body is properly constrained so that any movement is impossible and all support reactions and internal forces can be uniquely determined. German textbooks typically treat static determinacy formally in a separate section (e. g. Gross et al., 2011; Brommundt et al., 2007; Dankert and Dankert, 2013). Similar content is found in Hibbeler (2004) in section *5.7 Constraints for a Rigid Body* or embedded in Meriam and Kraige (2008) in section *3/3 Equilibrium Conditions*, although the term "static determinacy" does not appear in the table of contents. It can be concluded that the concept of static determinacy is taught in both contexts, but is more strongly emphasized in the German than in the US context.

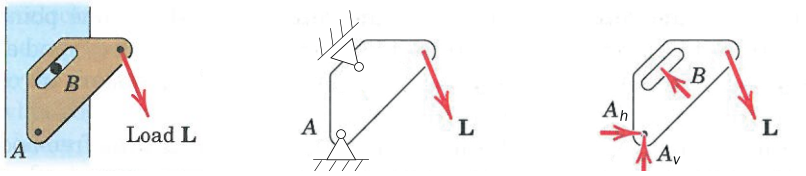
Conclusion: Static determinacy is more strongly emphasized in the German than in the US context.

MODELING Another noticeable difference relates to modeling systems. The step of modeling a real system such that the deduced model contains only relevant information is a key competency of engineers. "Students are able to illustrate relevant steps of modeling" is specifically mentioned as a learning goal of the inspected course (Hamburg University of Technology, 2019). Nevertheless, it seems to play a rather small role in traditional German introductory mechanics instruction. Gross et al. (2011) address and justify this instruction design in their introduction:

The learning goals are listed in Section 19.2.

> *"Since it is first of all a matter of learning the basic laws and their correct application, we will mostly exclude the question of modelling, which requires a lot of skill and experience. The mechanical analysis of idealized systems, in which the real technical starting point is sometimes no longer recognizable, is not, however, unrealistic gimmickry, but should enable the prospective engineer to later solve practical problems on their own with the help of theory."*
>
> (Gross et al., 2011, p. 2)[3]

3 German original: Da es zunächst auf das Erlernen der Grundgesetze und ihrer richtigen Anwendung ankommt, werden wir die Frage der Modellbildung, die viel Können und Erfahrung voraussetzt, meist ausklammern. Die mechanische Analyse idealisierter Systeme, in denen der reale technische Ausgangspunkt manchmal nicht mehr erkennbar ist, ist jedoch nicht wirklichkeitsfremde Spielerei, sondern sie soll den angehenden Ingenieur in die Lage versetzen, später praktische Probleme mit Hilfe der Theorie selbständig zu lösen.



(a) realistic (b) abstract (c) free-body diagram

Figure 16: Different degrees of abstraction. German instruction uses abstract support models (b), US instruction prefers more realistic representations (a) as starting point to deduce the free-body diagram (c). (Figure in (a) taken from Meriam and Kraige (2008))

See details in Section 11.1.1.

Some of the interviewed experts share this attitude and stated in the interviews that the step of modeling a real system is deliberately introduced only in later semesters because students should first learn how to solve statics problems. Others see the reason in the strict separation of the German engineering mechanics from the instruction on machine elements and design.

Generally, the use of photographs is more prevalent in US textbooks *"to show how engineers must first make an idealized model for analysis and then proceed to draw a free-body diagram of this model in order to apply the theory"* (Hibbeler, 2004, p. ix). US textbook authors Meriam and Kraige (2008) criticize the very abstract and theoretical approach of engineering mechanics instruction and argue that theory should be taught in the applied engineering context to better motivate student learning:

> *"There is a frequent tendency in the presentation of mechanics to use problems mainly as a vehicle to illustrate theory rather than to develop theory for the purpose of solving problems. When the first view is allowed to predominate, problems tend to become overly idealized and unrelated to engineering with the result that the exercise becomes dull, academic, and uninteresting. This approach deprives the student of valuable experience in formulating problems and thus of discovering the need for and meaning of theory. The second view provides by far the stronger motive for learning theory and leads to a better balance between theory and application. The crucial role played by interest and purpose in providing the strongest possible motive for learning cannot be overemphasized."*
>
> (Meriam and Kraige, 2008, pp. vii-viii)

Consequently, formulating engineering problems and thus modeling must be part of the student's activities. This difference in philosophy also affects the CATS and the mental process students need to pursue when solving the questions. The graphical representation of structural elements in statics (e. g. supports, joints, beams, trusses etc.), whether more realistic or more abstract, stands for certain rules and

assumptions which loads can act on or be exerted by these elements. If standard elements are learnt in connection with these rules and assumptions, the translation between the graphical representation and the rules happens automatically. Hibbeler (2004), for example, provides large tables linking a variety of support types to the possible reaction. One expert described this process as follows:

> *[Relates to CATS concept "Representing"]: "[...]. For me, the image [...] that stands behind it is this table with a certain support type and a corresponding image of the possible forces. [...] mentally I take a support away and replace the support with the equivalent forces. And that's what I'm actually testing with this item. So this has nothing to do with the axioms but rather with this copy-and-paste strategy, you know? So I have a certain image of a section with corresponding internal forces and I replace the support with them. [...]"*
>
> *[bezieht sich auf CATS Konzept "Representing"]: "[...]. Für mich ist das Bild, [...] das dahinter steht, ist diese Tabelle mit einem bestimmten Lagertyp und dem Bild der Schnittkräfte, das dazu gehört. [...] gedanklich nehm ich ein Lager weg und ersetze das Lager durch sein, durch das äquivalente Schnittbild. Und das ist das, was ich damit eigentlich abprüfe. Das hat also nichts mit den Axiomen zu tun sondern eher mit diesem ja Copy and Paste ist es so ein bisschen ne? Also ich habe ein bestimmtes Bild mit einem Schnitt, von Schnittgrößen die dazu gehören und die tu ich statt dem Lager da rein. [...]"*
>
> *E8 #00:33:04-7#*

US textbooks often prefer a more realistic representation as shown in Figure 16 (a), while most German textbooks traditionally use more abstract models as shown in Figure 16 (b). Even if German textbooks introduce supports at times as concrete implementations (e. g. Gross et al. (2011)), they are subsequently used only as abstract representations, with distinctions made only with respect to the number and type of constraints imposed by the support. Figure 16 illustrates possible consequences. When solving the CATS, US students are trained by instruction to directly infer Figure 16 (c) from Figure 16 (a). German students are not familiar with roller and pin-in-slot joints as used in the CATS. Instead, they are trained to infer (c) from (b) and thus may have to take an additional mental step from (a) to (b) to translate the problem into the familiar abstract language.

CONCLUSION This additional step from a realistic to an abstract representation suggests that CATS results from the US and from Germany may not be directly comparable and that German students may require more time. Furthermore, the German CATS may test one further construct, i. e. whether students are able to take that additional step.

11.1.4 *Summary: Content*

To determine whether the CATS addresses appropriate content to measure the construct *conceptual understanding of statics*, expert interviews were conducted, the documented course content of the investigated course was consulted, and German and US textbooks were compared.

The CATS does not go beyond the content generally addressed in introductory mechanics courses in Germany. The limited scope allows for only nine concepts to be addressed. For example, the complex concept cluster of FBDs involves several concepts, of which only one is directly probed by the CATS. However, the concepts selected by the CATS developers were confirmed to be essential, with a possible exception of the Roller concept (and analogously the Pin-in-slot concept). In general, German students do not typically learn about specific implementations of support types. Instead, the concept of static determinacy is of higher importance in the German context. The German CATS may furthermore test an additional construct: the students' ability to deduce abstract models from realistic representations. Consequences are that German students may require more time than US students and that German test results should not be compared to US test results.

11.2 CRITERION VALIDITY

Criterion validity is established by correlating the CATS total score with performance on course examinations. Furthermore, similarities between the students of the US and the German context are established by comparing the most attractive distractors.

11.2.1 *Correlation with exams*

Scatter plots of the matched subsamples for which both, CATS and exam data, are available indicate a moderate correlation between the two instruments in both cohorts (Figure 17 (c) and (iii)). The correlations between the CATS as post-test to statics instruction and the statics exams (midterm and final) were both found to be $r = 0.5 \pm 0.1$,

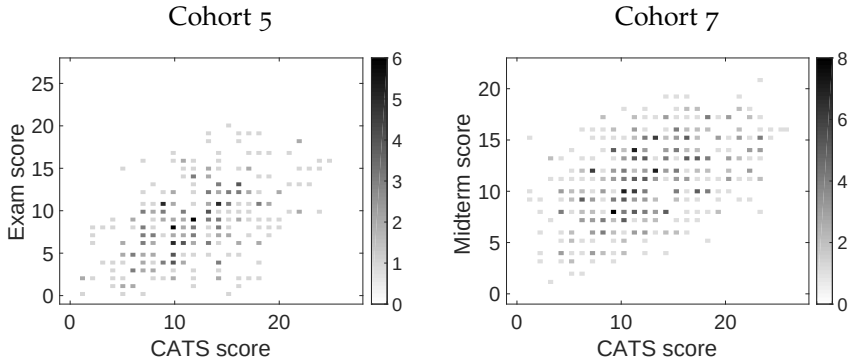


Figure 17: Scatter plots of CATS and exam scores for two cohorts show a moderate correlation in each case. Grayscale indicates measured frequency of score combination.

which is comparable to results from the US context (Steif and Hansen, 2006a).

It may be surprising that the correlations are the same for midterms and finals. The CATS was designed to address basic and essential concepts, which are expected to be introduced halfway through the curriculum. So the midterm likely contains a larger share of the CATS concepts than the final exam and could therefore also be expected to correlate stronger. One possible explanation is that the basic concepts are also essential for the more advanced content on the final exam (which would justify them to be categorized as basic). Evidence for this effect was observed by Steif and Hansen (2006a) (→ p. 62).

11.2.2 *Analysis of distractors*

A multiple-choice test instrument is only as good as its distractors. If the correct response option is evident without the need to apply the understanding of the concept to be tested, the instrument does not measure what it claims to measure. With similar distractors on the items in a concept category, the CATS is designed to serve as a diagnostic tool for misconceptions with repeated measurements for increased precision. Even though it is not used for this purpose in this dissertation, evidence that misconceptions can be reliably detected because distractors are chosen consistently, adds to the collection of evidence that the total test score can be interpreted as intended. Furthermore, similar patterns of distractor attractiveness in German and US contexts indicate that the test may be suitable for both populations.

The logic behind the distractors was already discussed in Section 6.1. Here, the data is examined with regards to the most attractive distractors on each item or concept category for students of different ability levels by looking at the items' IRCs as introduced in Section 9.3.2. The findings are summarized here and compared to data from the US

context found in literature. The detailed analysis can be found in Appendix F.

DRAWING FORCES CONCEPT The results of the single items draw a coherent picture. Most students who miss these items seem to be confused about whether or not to include internal forces in a free-body-diagram (IntF). The second largest difficulty is understanding that forces acting on external bodies must not be drawn as they are accounted for in the contact forces acting directly on the modeled system (W+F). These findings are in line with the results published by Steif and Hansen (2006a) and Román et al. (2010a) (see Table 21 in Appendix F).

See description of distractors in Section 6.1.

NEWTON'S THIRD LAW CONCEPT In accordance with Steif and Hansen (2006a) and Román et al. (2010a), the dominant distractor is 2Force. Instead of applying Newton's Third Law (and the equilibrium condition for the pin), students missing these items tend to imagine the forces along the members. Possibly, the systems are misinterpreted as trusses that can only exert longitudinal forces along their members.

This misinterpretation could be observed in the student interviews (→ Section 11.3.3.2).

Although the Newton's Third Law category did not yet exist at the time, Steif and Hansen (2006a) report that "[i]n general, there is a strong tendency to assume that forces always act parallel to elongated members". These findings are thus also in line with the results published by Steif and Hansen (2006a) and Román et al. (2010a).

STATIC EQUIVALENCE CONCEPT In accordance with Steif and Hansen (2006a) and Román et al. (2010a), the misconception that a single force and a couple are interchangeable (M=F) is dominant across all items with the exception of item 7, where distractor CentrM (d) is most attractive to both, the very low and very high scoring students, while M=F (b) is most attractive only in the medium score range.

ROLLER JOINT CONCEPT Judging by the student subscores on the concepts, the concept of rollers seems to be one of the easier concepts. Low scoring students tend to choose the distractor resembling the misconception that the force exerted by a roller mounted on a low-friction bearing acts along the arm supporting the roller (2Force), whereas high scoring students tend to choose ($\arctan(\mu)$).

Steif and Hansen (2006a) and Román et al. (2010a) do not discriminate between high and low scoring students, but they report the most attractive distractor to be 2Force. Román et al. (2010a) furthermore reports the second-most attractive distractor to be $\arctan(\mu)$ which complies with the results presented here.

PIN-IN-SLOT JOINT CONCEPT Nearly linear IRCs (see Figure 54 in Appendix F) of the correct responses to items 3 and 12 indicate that these discriminate no better than the total score on the test (Morris et al., 2006), i. e. they do not add value to the discriminatory power of the total test in any specific score region (but they might add value to the concept category). Item 21 shows a neat s-curve with slightly better discrimination.

Throughout the pin-in-slot items, (Moment) is the dominant distractor, which involves the correct direction of the force, but also a couple acting on the frictionless pin. The results are in accordance with Román et al. (2010a). Steif and Hansen (2006a) do not report data on this concept.

The response that the force on the pin can act along the slot (Motion, distractor (d) in all items), which might be chosen if a possible motion and force are confused, is quite common on item 3, while it hardly shows on both the other items. One difference between items 3 and 12 is the orientation of the force along the slot. In item 12, the force along the slot has a large component in the *same* direction as the applied force, while in item 3, the force along the slot has a large component that *opposes* the applied force. (In this respect, item 21 is the same as item 3, but is difficult to compare because of the large number of blank responses.) The opposing direction seems to be slightly more attractive, but the reason for this preference is unclear, especially because it contradicts the hypothesized confusion of motion and force. Further research would be required to make definite statements.

NEGLIGIBLE FRICTION CONCEPT The concept seems to be generally difficult for students. The items all ask the students to assess two cases respectively as possible or impossible, a design which only has four "natural" distractors. The fifth distractor ("not enough information") was added only for structural reasons, which shows in the data as very low distractor attractiveness.

Low scoring students tend to assess both cases incorrectly while high scoring students, who yet miss the respective item, tend to do better in the sense that they assess one case correctly. This adds to the evidence that the test score scale is a valid measure of understanding of statics. The results are in accordance with Román et al. (2010a), except for item 25, where the US students mostly select distractor (a) (both possible). Steif and Hansen (2006a) do not report data on the NeglFric concept.

REPRESENTATIONS CONCEPT Judging by the low selection rate of distractors with incorrect numbers of possible forces, students seem to be rather confident in determining how many different *forces* a certain support can generally exert. Difficulties arise with respect to *moments*, which fits the overall picture that moments are a difficult

Misinterpretation of the point of interest (→ p. 134).

concept. The distractor analysis furthermore supports observations made in the expert interviews that item 5 should be reworded.

On two items (14 and 23), the average rankings of the distractors found here and published by Román et al. (2010a) differ (compare data shown in Appendix F, Table 20 and Table 21). The German students believe more frequently that the rope can exert a moment in addition to the tension force than believing that the force acts at an arbitrary angle. The US students hold both ideas equally often. An even stronger discrepancy is found in the rankings of the distractors to item 23, but this may be only due to the very small number of students even selecting distractors. Steif and Hansen (2006a) do not report any data on the Representations concept.

LIMIT OF FRICTION CONCEPT Mistaking the limit of the friction force as the actual friction force in the static case is the most common distractor (MuN) followed by the difference between the limit of friction and the driving force (F-MuN). These results are in line with Román et al. (2010a) and Steif and Hansen (2006a).

EQUILIBRIUM CONCEPT The Equilibrium concept cluster is the most remarkable one as it includes the most problematic item 20 and the easiest item 11. Furthermore it has two of the later items in the test, which shows in the large amount of blank responses, especially in the low score ranges.

Steif and Hansen (2006a) report data on an older version of item 11, where the most dominant distractor resembles (b), a moment can balance a force (compares to M=F from Static equivalence). The subsequent adaptations made to the item obviously reduced the selection rate of this distractor, as Román et al. (2010a) also report distractor (e) to dominate (force at 20 degrees plus a moment), as it was found here. Steif and Hansen (2006a) report on item 20 that only 11 % responded correctly, while 70 % stated that body (I) (imbalance of forces) could be in equilibrium and 53 % accepted that body (II) (imbalance of moments) could be in equilibrium and Román et al. (2010a) report these numbers as 16 % (correct), 54 % (imbalance of forces accepted), and 57 % (imbalance of moments accepted). Here, the respective numbers are similar with 11 % (correct), 58 % (imbalance of forces accepted), and 51 % (imbalance of moments accepted), if blank responses are omitted. The two most attractive distractors on item 26 match the ones reported by Román et al. (2010b), although with reversed ranks.

11.2.3 *Summary: Criterion*

Calculating Pearson's r of CATS scores with examinations revealed a moderate positive correlation of $r = 0.5 \pm 0.1$ as expected. Similar values have been previously reported for the US context.

Further similarities between the German and US contexts have been established by the distractor analysis which was conducted using IRCs, a method that allows to examine the selection rate of each distractor in dependence of the students' total score. Thereby, differences in distractor attractiveness could be revealed in some items for low- vs. high-scoring students, which have not been previously reported to the best knowledge of the author. This technique also allowed to confirm the theoretical ranking of distractors (where applicable), i. e. that higher scoring students tend to select "less incorrect" distractors than low scoring students. On all concepts, the ranking of the distractors in terms of average selection rate largely matched results reported in literature form the US context.

These results are strong evidence that CATS test scores are likewise interpretable in the German higher education context and that the instrument may even be used as a diagnostic tool for common misconceptions in statics.

11.3 CONSTRUCT VALIDITY

Construct validity is the most difficult aspect of validity to establish. In this section, the results of the mixed-methods approach introduced in Chapter 9 are presented. One of the main threats to construct validity, interpretability of the items, is investigated by means of qualitative data from expert and student interviews as well as inspection of the quality of translation. Subsequently, following the framework proposed by Jorion et al. (2015), quantitative evidence that the data fits models from test theoretical frameworks and factor analysis is presented. Finally, the results are compared to literature from the US context.

11.3.1 *Expert interviews (interpretability)*

The responses to two questions on the expert interviews mainly address interpretability of the items. These are question 2, regarding any perceived weaknesses of the test, and question 7, regarding the abstraction level of the items.

Did you notice any weaknesses in the test or individual items?

The experts noticed a variety of aspects which they perceive as weaknesses. First, general criticism is presented, followed by remarks relating to conventions. Finally, otherwise unclear or confusing aspects are addressed.

1. General criticism
    a) Instructions

   b) Precision of items

2. Conventions
   a) Friction terminology
   b) Forces in components vs. forces as two-dimensional vectors
   c) Weight force notation

3. Otherwise unclear/confusing aspects
   a) No (weight) force on Roller items
   b) Block support on item 2
   c) Point of interest on item 5
   d) Arm support on item 8
   e) Incomplete "free-body diagrams" on Limit of Friction items

GENERAL CRITICISM The instructions at the beginning of the test were introduced by the developers to state the general assumptions applicable to the entire test. Some experts mentioned that stating them once at the beginning is not enough, as students would not remember. Experts noticed that sometimes instructions are repeated on individual items, but not consistently. These experts then suggested that the instructions should be repeated on every item they apply to.

> *"Yes, it is reoccurring that in some places the term 'frictionless' is mentioned and in others not. [...] and I find that inconsistent. [...] I was confused and then I turned back again and thought: 'Well. What applies now?'"*
>
> *E8: "Ja das ist immer wieder, an einigen Stellen steht ein 'reibungsfrei' dabei und an anderen nicht. [...] und das finde ich dann inkonsistent. [...] ich war verwirrt und hab dann wieder nach vorne geblättert und hab gedacht: 'Ja. Was gilt denn nun?'"*
>
> *E8 #00:43:09-8# ff.*

Similarly, items were criticized as not being precise enough. In this case, adding descriptive text was often suggested as an improvement.

> *E1: [refers to item 3] "[...] should be guided without clearance. Is that actually noted here somewhere?"*
>
> *I: "No."*
>
> *E1: "[...] We would phrase it like this, so... [writes] 'At point A the lever is mounted with a pin in a slotted hole such that it is free of play and friction.'*

E1: [bezugnehmend auf Item 3] "[...] soll ja spielfrei geführt sein. Steht das eigentlich hier irgendwo?"

I: "Nee."

E1: "[...] Wir würden jetzt sowas hinschreiben [...] [schreibt] 'Am Punkt A ist der Hebel über einen Stift spiel- und reibungsfrei in einem Langloch geführt.'"

E1 #00:21:51-7#

At the same time, the amount of text was mentioned as one cause of the time problem.

"I always think it's pretty much to read in the short time."

"Ich finde immer es ist relativ viel zum Lesen für die kurze Zeit."

E1 #00:46:25-3#

Only one expert suggested a different solution. He noticed that text describing the images is generally redundant information and could be removed.

"It is so overdone with text and then image. I think one should understand a free body diagram or an image or a drawing much better without the need to really have both [...]".

"Es ist so übertrieben mit Text und danach Bild. Ich glaube man soll ein Freikörperbild oder ein Bild oder eine Zeichnung viel besser verstehen, ohne dass man das wirklich so doppelt machen muss [...]."

E11 #00:06:39-2# ff.

There seems to be a conflict of requirements which can be illustrated by a variant of the quality triangle, a concept which is frequently used in project management (see Figure 18). Ideally, the measurement should not cost any time while measuring the entire course content very precisely. This is of course impossible, because there are only two degrees of freedom among the set of three variables *scope*, *precision*, and *time*. The quality of the test measurements is of course affected by the scope of the test. In order to measure conceptual understanding of statics, the test should address as many course-relevant concepts as possible. Likewise, the quality of the test is affected by the precision of the measurement, which can be increased by two approaches: (1) increasing the number of items to repeatedly measure the same concept with slightly different items (as done on

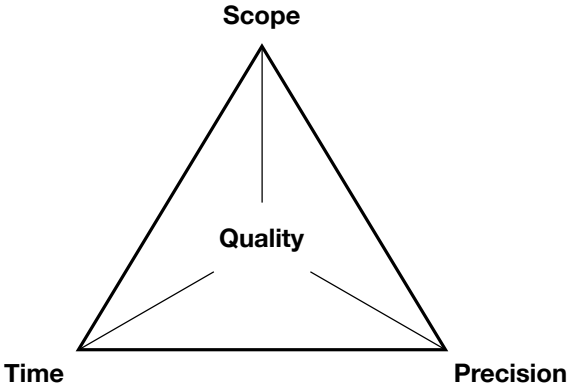


Figure 18: Quality triangle

the CATS) to average out any measurement error or (2) increase interpretability by adding descriptive text and repeat the basic rules and assumptions. The latter is the preferred strategy among the experts. Both requirements, wide scope and high precision, inevitably lead to high costs in terms of test time. Usually, test time is limited, therefore cutbacks must be accepted in scope and/or precision.

CONVENTIONS: FRICTION TERMINOLOGY Semantically, some experts generally criticized the use of the terms "Gleitreibung" (kinetic friction) and "Haftreibung" (static friction), as used for example on the Limit of Friction items. Instead, they prefer the terms and symbols "Reibung" (μ) and "Haftung" (μ_0) to emphasize the difference. In one case, the different terminology even led to a severe misunderstanding on the Roller items. E5 was convinced that the two statements on friction (frictionless roller bearing and non-zero coefficient of friction) contradict each other.

> *"[...] It says there: [...] 'There is friction between the rollers and the body, with a coefficient of friction of* 0.6*'.* (author's remark: the wording in the item is in fact different.) *In my opinion this is a clear contradiction to the previous statement that the rollers' bearings are frictionless."*
>
> *"[...] Da steht da: [...] 'Zwischen den Rollen und dem Körper gibt es Reibung, mit einem Reibungskoeffizienten* 0,6*'* (Anmerkung der Autorin: Item falsch zitiert.)*. Das ist in meinen Augen ein klarer Widerspruch zur vorigen Aussage, dass die Rollen reibungsfrei gelagert sind."*
>
> *E5 #00:23:34-2#*

At first, it seemed as if he missed that the item stem only specifies a coefficient of friction and does *not* claim that there is a friction *force* between the roller and the block. After closer inspection of the transcript, it turned out to be a misunderstanding of the term "Reibung"

(friction). By the following statement, he acknowledges that there is "Haftung" (static friction) between the roller and the body, but no "Reibung" (kinetic friction) due to the frictionless roller bearing.

> *"The [rollers] are allowed to rotate, the bearings are frictionless, then there's [static] friction between roller and [body], but the rollers rotate. They can only transmit normal forces, there's no [kinetic] friction, okay?"*
>
> *"Die [Rollen] sind drehbar gelagert, hier in den Lagern reibungsfrei, dann gibt's hier eine Haftung zwischen Rolle und [Körper], aber die Rollen drehen sich ja. Die können nur Normalkräfte übertragen, da gibts keine Reibung, ja?"*
>
> *E5 #00:24:13-4# ff.*

Both terminologies are used in Germany. A choice for using one over the other could thus lead to confusion in the different traditions of German courses, either way. It is unclear whether the same problem or other problems could occur if the terminology on the CATS was changed to "Haftung/Reibung". The student interviews cannot provide more insight on this issue as all students were drawn from the same course, using the terminology also used on the CATS.

CONVENTIONS: FORCES IN COMPONENTS VS. FORCES AS TWO-DIMENSIONAL VECTORS Some experts expressed concern about using forces as two-dimensional vectors rather than horizontal and vertical components.

Two experts (E1 and E7) stated that their students are not supposed to think in or use complete two-dimensional vectors but always components, another two (E5 and E6) were indifferent, and one (E8) would actually prefer to emphasize that forces are vectors:

Consequences of using only components were observed in the student interviews on item 25 (→ p. 150).

> *"[...] from my point of view the point is that the students learn how to translate a certain system into computational software language. With this perspective the view of a force as a vector is much more important [...]."*
>
> *"[...] eigentlich geht es ja darum, dass die Studierenden lernen wie sie ein bestimmtes System in den Computer bringen, aus meiner Sicht. Und dann ist eigentlich die Sichtweise einer Kraft als Vektor viel wichtiger [...]."*
>
> *E8 #00:35:43-3#*

However, in case the two-dimensional force vector points into a "special" direction (e. g. horizontal, along structure parts etc.) under otherwise arbitrary conditions, most experts see this as problematic.

The symptoms were visible on item 4, which tests students on their understanding of Newton's Third Law. The question asks for the directions of the forces from a connecting pin on connected frame parts. The response options show force vectors acting in various directions, including along the frame parts and perpendicular to them. Only one option features a force pair pointing in opposite directions in accordance with Newton's Third Law. The developers chose to draw these forces horizontally. One expert explicitly stated that the special direction of the force did not bother him (E8 #00:31:21-7# f.), but five experts (E1, E3, E5, E6, E7) criticized that the correct response is a special case and should not be used in such a situation which stands for arbitrary loads.

> *"Now [...] the proposed solution is such that it is very horizontally oriented and the [students] then say: 'And what about the possible vertical part? [...]' Our [students] might stumble over it."*
>
> *"Jetzt ist [...] der Lösungsvorschlag so, dass der sehr horizontal orientiert ist und die [Studierenden] dann sagen: 'Und was ist jetzt mit dem möglichen vertikalen Anteil? [...].' Da würden unsere [Studierenden] vielleicht drüber stolpern."*
>
> *E3 #00:22:42-7#*

> *"[T]his is a special case and I would never sketch a special case, [...] because the directions of the [external loads] are completely arbitrary, [...] then the directions of the [inter-body reactions] will probably be arbitrary, too."*
>
> *"[D]as hier ist ein Spezialfall und ich würde nie einen Spezialfall, sozusagen hier skizzieren, [...] weil die [äußeren Belastungen] völlig beliebige Richtungen haben, [...] dann werden die [Zwischenreaktionen] wahrscheinlich auch beliebige Richtungen haben."*
>
> *E6 #00:31:57-2# ff.*

Experts thus anticipate that even students who understood the concept of Newton's Third Law may hesitate to choose the correct response because they know that the pin is able to exert forces in not only one but two dimensions. Students who are not familiar with arrows as indicating complete two-dimensional force vectors from instruction are especially at risk.

CONVENTIONS: WEIGHT FORCE NOTATION Ideally, any concept which is not the focus of the item should not be addressed by it. Giving weight forces instead of masses has the advantage that it eliminates the unnecessary step of calculating the weight forces from the masses. The notation of the block's weight forces as scalars in the

Limit of Friction items was often criticized as incorrect, but also acknowledged to be well interpretable:

> *"[...] I can't assign 60 Newton to a body [...]. I can assign a weight to it, [...] for a force I have to sketch in a vector. [...] But [...], I don't want to be pedantic either, I understand what they mean."*
>
> *"[...] ich kann nicht an 'nen Körper 60 Newton dranschreiben [...]. Ich kann ein Gewicht dranschreiben,[...] für eine Kraft muss ich dann einen Vektor einzeichnen. [...] Aber [...], ich will auch nicht pedantisch da sein, ich versteh was gemeint ist."*
>
> *E8 #00:43:46-4#*

In two of the Limit of Friction items (6 and 13), the item stems describe the blocks as lying on a table so that the direction of the weight forces must be concluded as pointing towards the table. Alternative situations of, for example, tilted tables are technically possible but very artificial. The direction of the weight force is thus assumed to be clear. Item 22, in which a block is clamped by two other blocks, is indeed less clear on the direction of the weight force. Here, drawing the weight force as a vector would not only be reasonable, but also easily implemented because of the available space below the 6 N block.

Students were interviewed on items 6 and 13. The items were interpreted correctly in all six cases.

POSSIBLY CONFUSING: NO (WEIGHT) FORCE ON ROLLER ITEMS The following describes an error found in two of the Roller items. The instructions state that "gravity forces are negligible unless on bodies with marked weights". It was noted by E1 that applying this assumption to Roller items 1 and 10 is problematic. Neither the block in item 1 nor the platform in item 10 is marked with a weight or subjected to a force which could counteract the force exerted by the roller. If static equilibrium is assumed (as given by the instructions), the force by the roller on the block/platform must be zero and thus does not have a direction. Technically, none of the response options are hence correct. While this error should definitely be corrected in future versions of the test, the consequences are minor. When prompted, many experts (and students) reported that they did not notice this error. In case the weight was *not* intuitively assumed, none of the distractors are more plausible than the expected response. Thus the test results are not expected to be biased. The student interviews (Section 11.3.3) support this expectation.

POSSIBLY CONFUSING: BLOCK SUPPORT ON ITEM 2 Item 2 shows a system of blocks connected by cords. Block 4 rests on blocks 5 and 6 which are suspended by one cord each. Two experts (E1 and E7)

noted that their focus was drawn to whether block 4 could be supported as illustrated or would slip through. One of them assumed that this would likewise be noted by the students and distract them from the desired task. This could not be observed in the student interviews. None of the four students even addressed block 4.

POSSIBLY CONFUSING: POINT OF INTEREST ON ITEM 5 Item 5 asks for possible support reactions on a pin support. In the German translation, the point of interest is referred to as "the support in the lower left corner of the body"[4]. Three experts (E5, E6, and E7) misinterpreted the point of interest to be the lower left support of the entire *system*, which refers to the attachment of the cord[5] to the floor or wall. This misinterpretation is most likely caused by not reading the text of the item stem carefully enough. The three experts all suggested to label the point of interest and to refer to it by that label in the text in order to avoid confusion (E5 #00:38:13-5#, E6 #00:16:03-2#, and E7 #00:20:02-0#).

The misinterpretation of the point of interest would lead to distractor (e). As shown in Figure 19, this distractor is equally attractive with a selection rate of about 10 % for all students, independent of their score. In the higher score range, it is the most frequently chosen distractor. This pattern is rather uncommon among the early items, and it suggests that this confusion of the point of interest also occurs among students, but it does not seem to happen at an alarming frequency. It did not occur in any of the three student interviews involving this item. Nevertheless, avoiding this confusion by introducing a label would be an easy adaptation that would improve the quality of the item in terms of interpretability.

POSSIBLY CONFUSING: ARM SUPPORT ON ITEM 8 Item 8 shows an arm driven by a torque in frictionless contact with a rocker. Two experts (E5 #00:39:25-1# and E8 in notes) explicitly noted that it is not clear how the arm in item 8 is supported. This question briefly came up in one student interview, but the student quickly realized that it was irrelevant.

POSSIBLY CONFUSING: INCOMPLETE FREE-BODY DIAGRAMS ON LIMIT OF FRICTION ITEMS The response options on the Limit of Friction items were interpreted as incomplete free-body diagrams, a representation which was heavily criticized.

> *"Yes, it bothers me very much that they offer incomplete free body diagrams here."*

4 "Gegenstand dieser Aufgabe sind die Lagerreaktionen, die das Lager in der linken unteren Ecke des Körpers auf diesen ausübt."

5 Furthermore, the cord was often misinterpreted as a rod, a detail without consequences.

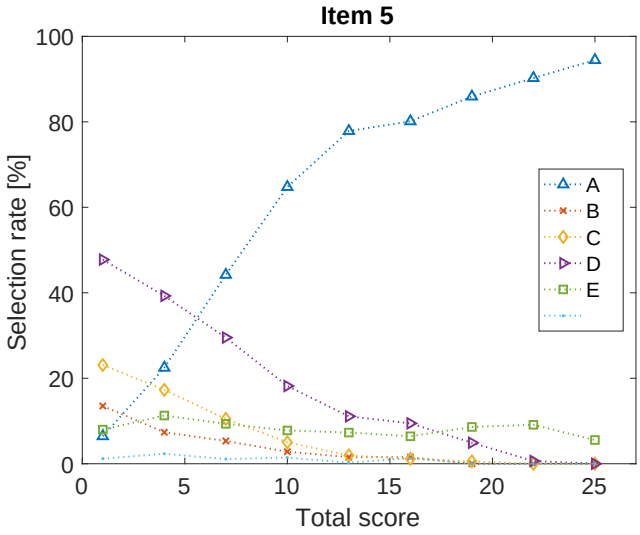


Figure 19: Item response curves (Morris et al., 2006) of item 5. Distribution of the responses (including blanks) depending on total test score. Option A is correct. Distractor E would be the logical response if the point of interest was misinterpreted.

> *"Ja, da stört mich sehr daran, dass sie unvollständige Freikörperbilder hier anbieten."*
>
> *E5 #00:38:49-6#*

> *"I find that problematic from a didactical point of view [...]. Because I always strongly emphasize that a free body diagram must be complete when I draw one.*
>
> *"Das finde ich aus Sicht der Didaktik [...] problematisch. Weil ich immer großen Wert darauf lege, wenn ich ein Freikörperbild anzeichne, dass es komplett und vollständig ist."*
>
> *E8 #00:14:53-5#*

The developers' motivation for these illustrations is probably to clarify which force the question asks for and not to show incomplete FBDs. The latter could be indeed problematic, unless the task is explicitly designed to complete the FBD. Implicitly, solving this CATS item does involve completing the "free-body diagrams" in the response options and applying the equilibrium condition for the horizontal forces to answer the question. Therefore, the interpretation of the response options as incomplete FBDs would not necessarily lead to an incorrect answer.

How do you assess the level of abstraction?

Most experts agree that the level of abstraction is different from the one used in their teaching but do not see this as problematic.

> *"Well, I don't have much of a problem with that. But it's true, we often [sketch] the exercises, let's say, more abstract, not so detailed [...]."*
>
> *"Also da habe ich wenig Probleme damit, ja. Aber es ist richtig, wir [skizzieren] häufig die Aufgaben, sagen wir mal, abstrakter, nicht so detailgetreu [...]."*
>
> *E5 #01:08:01-7# ff.*

> *"[...] so the idea is of course that we want to train the students to see abstraction in what we do, but to approach it so dogmatically and say: A platform can only be a line and a support can only be these triangles, circles and such, [...] I find that a bit boring, I have to say."*
>
> *"[...] also die Idee ist natürlich schon, dass wir die Studierenden ja auch dazu trainieren wollen also die Abstraktion zu sehen in dem was wir da machen, aber da jetzt so dogmatisch ranzugehen und zu sagen: Eine Plattform darf nur ein Strich sein und ein Lager darf nur, also diese Dreiecke, Kreise und sowas, [...] das finde ich jetzt ein bisschen langweilig, muss ich sagen."*
>
> *E8 #00:24:21-8#*

> *"[...] I use other symbols for these, for these pin supports for example, but I think one can interpret this as well. [...]"*
>
> *"[...] Ich benutze dann irgendwie andere Symbole für diese, für diese Festlager zum Beispiel aber ich finde das kann man auch erkennen. [...]"*
>
> *E9 #00:35:02-8#*

E2 states that

> *"[...] [the CATS] is not badly developed, because it often achieves a balance between unnecessary details and relevant elements."*
>
> *" [...] [der CATS] ist schon auch nicht schlecht entwickelt, weil es oft die Balance eben zwischen unnötigen Details und relevanten Elementen trifft."*
>
> *E2 #00:17:30-7#*

He furthermore notices that both the abstract and realistic representations are associated with (possibly different) assumptions which must be correctly interpreted by the students and made explicit by the instructors and the test in any case (E2 #00:12:15-5#). When the interviewer mentioned that pre-instruction students can in general interpret the graphical representations and were found to struggle only

with technical terms such as moments and free-body diagrams (as reported in Chapter 8), E2 concluded that it makes sense to operate with the more intuitively interpretable representations such as rollers and not with abstract support symbols to focus on the essential concepts such as forces and moments (E2 #00:19:03-4#). Others believe that only abstract representations allow to focus on the essential statics problem.

> *"[...] Here, a support is not only sketched roughly, but the roller itself or something more concrete. That, for example, bothers me again because the students.....well, their idea of the implementation is too specific. Statics lives on... mechanics lives on extreme abstraction. [...] [In the determination of the unknown parameters] I find it disturbing when pictures are presented, which are supposed to all look a bit more like practice and are supposed to be more concrete and vivid. [...]"*
>
> *"[...] Da wird dann nicht nur so ein Lager grob gezeigt, sondern wirklich dann eine Rolle oder irgendwas konkreteres. Nur das stört mich zum Beispiel schon wieder, weil da die Studenten.....ja, ich sag mal, eine zu klare Vorstellung von der Umsetzung haben. Statik lebt ja davon... Mechanik lebt davon, extrem weit zu abstrahieren. [...] [Bei der Bestimmung gesuchter Größen] finde ich es störend wenn dann aber Bilder kommen, die alle so ein bisschen mehr nach Praxis aussehen sollen und gegenständlicher und anschaulicher sein sollen. [...]"*
>
> *E4 #00:04:02-4#*

According to E4, first-year students may be confused by the realistic representations as they are not expected to create abstract models from real systems in the course.

This difference in abstraction levels was also noted in comparing US American to German textbooks (→ Section 11.1.3).

> *I: "[...] Later, an engineer has to do exactly that [...] deriving abstract models from real-life systems"*
>
> *E4: "But not in the first semester. [...] We take this step of abstraction for the students in the first semester."*
>
> *I: "[...] ein Ingenieur später muss ja diese Arbeit auch leisten [...] von realen Systemen zum Modell [...] abstrahieren."*
>
> *E4: "Aber nicht im ersten Semester. [...] Wir fangen halt so an, dass wir diesen Abstraktionsschritt im erstem Semester vorwegnehmen."*
>
> *E4 #00:05:56-5# ff.*

E7 agrees with E4 that the CATS requires modeling competence, which is *"an aspect that is being tested a little here [...]"* (E7 #00:03:45-2#) but, contrary to E4, he emphasizes modeling as a basic skill which is included in his course, and which can be tested by the CATS in a very early phase, *"perhaps after four weeks"* (ibid). In this respect, he also criticizes the scope of the CATS as not being comprising enough of the course content to infer a measure for expert-like understanding of statics from the test score.

SUMMARY EXPERT INTERVIEWS The interviews disclosed a conflict of interest between scope of the test, time on the test and precision of the items. Experts suggested to be more precise by adding text to the items. The latter aligns with the results from the translation analysis (presented below). Experts named a few incompatibilities with the conventions they use and pointed out further unclear or confusing aspects, most of which were not shared by the students according to the interview data (presented below). Although the experts disagree on whether or not system modeling should be an intended learning outcome for a first-year course, the items were evaluated to be interpretable with regard to the level of abstraction. This could be confirmed by the student interviews (presented below).

11.3.2 *Translation analysis*

The translation of the test from English to German resulted in several deviations of different kinds. The translation analysis revealed three categories of motivations for adaptations made in the process of translating the CATS:

1. Precision: The translated version was making an increased effort to formulate the items precisely and make them "bullet-proof".
2. Terms and notations: Unfamiliar terms and notations were paraphrased and explained.
3. Conventions: Conventions were considered.

Table 4 lists some examples of systematic deviations found in the translation. They will be elaborated in the following paragraphs.

11.3.2.1 *Motivation category 1 - Precision*

One aspect was repeatedly found in all items: If not already explicitly stated in the original version, the phrasing was modified to clarify

Table 4: Examples for systematic deviations in the translated version from the original.

Items	Description of deviations	Motivation category
Drawing forces [2, 9, 15]	Cords are referred to in the text by their names	1
	Forces are renamed ($W \rightarrow$ G and T $\rightarrow$ S), and their nomenclature is explained.	3 2
Pin-in-slot [3, 12, 21]	Question mentions couple in one response option	1
Items involving arrows [11, 25]	Additional sentence stating that the arrows stand for forces and couples	1
Representation of loads [5, 14, 23]	Avoided term "partial free body diagram", instead, partial nature of the diagrams is mentioned separately as a note	2
Items involving the term "load(s)" [various]	The literal translation "Lasten" was avoided. Instead, the term "forces (and moments)" was used.	2
Newton's Third Law [4, 16, 24]	Omitted "Forces act in the senses shown."	3
Items involving arbitrary forces [4, 5, 14, 23, 24]	Explicitly stated that shown forces indicate only possible, but not actually acting forces.	3

that arrows stand for forces and moments in all items. These modifications are clearly driven by the first category. In most cases, the modifications are subtle (e. g. "Which arrows indicate the direction of the force..." instead of "What is the direction of the force..."). Items 11 and 25, however, are extreme examples. Although not the first items on the test with arrows, the translation of these items contains an additional sentence stating that arrows indicate the directions of forces and couples. The need for this additional hint on specifically these items is not immediately apparent. The insights from the student interviews, however, suggest that students indeed tend to misinterpret the arrows on item 25 as directions of possible reactions in the form of motion instead of force and couple - despite the additional statement given. This phenomenon may justify the need for giving the additional hint on this item, but at the same time proves it to not be effective for all students.

A suggestion for improvement of item 25 can be found in Appendix G.

The question on all "Pin-in-slot joint"-items serves as another example for motivation category 1. The question in the original version *"What is the direction of the force exerted by the slot on the pin A?"* ignores that one distractor also includes a couple (see Figure 20). In the translated version, the couple is mentioned in the question, resulting in a much more complicated sentence, also due to the more complicated German grammar. Furthermore, while the english phrasing of the question is the same in all three items, the German phrasing is different in every item (gray text translates back to English):

> *Welche Pfeile geben die tatsächliche Richtung (bzw. den Drehsinn) der Kraft (bzw. des Moments) an, welche die Aussparung in dem schrägen Bauteil auf den Stift A ausübt?*
>
> *Which arrows indicate the actual direction (respectively, the actual rotary direction) of the force (respectively, the moment), which the slot in the inclined element exerts on the pin A?*
>
> *Item 3*

> *Welche Richtung hat die Kraft (bzw. welchen Drehsinn hat das Moment), die von der Aussparung auf den Stift im Punkt A ausgeübt wird (bzw. werden), wenn sich der Mechanismus in der im Bild dargestellten Anordnung befindet?*
>
> *Which direction does the force (respectively, which rotary direction does the moment) have, that is (respectively, are) exerted by the pen in point A, if the mechanism is oriented as shown in the illustration?*
>
> *Item 12*

> *Welche Richtung hat die Kraft (bzw. welchen Drehsinn hat das Moment), die (das) von der Aussparung auf den Stift im Punkt A ausgeübt wird (werden), wenn sich der Mechanismus in der im Bild dargestellten Anordnung befindet?*

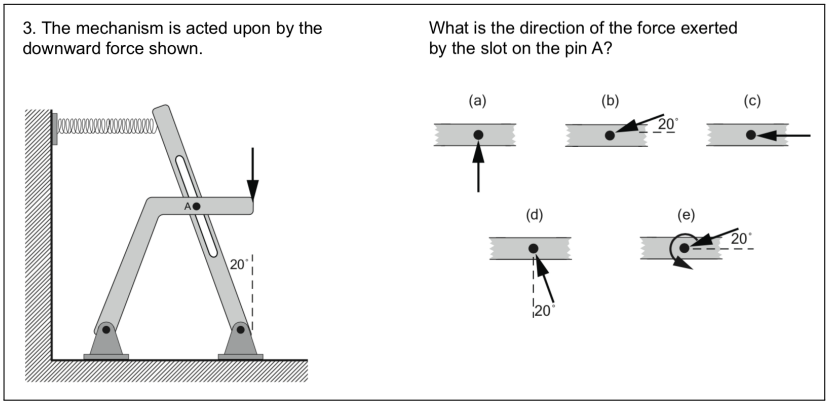


Figure 20: Example for Pin-in-slot item.

Which direction does the force (respectively, which rotary direction does the moment) have, that (that[a]) is (are) exerted by the pen in point A, if the mechanism is oriented as shown in the illustration?

Item 21

a attributed to the different genders of the German nouns "force" and "moment"

In addition to the complicated phrasing, the following flaws were identified:

1. Asking for the *direction* of a moment is not meaningful if there is only one option given.
2. To account for the moment in option (e), the phrasing on item 3 asks for arrows in the plural ("Which arrows..."), which may falsely suggest that multiple responses are correct.
3. It is emphasized on items 12 and 21 that all parts of the mechanism are oriented as shown, but not on item 3. This difference may falsely suggest that the orientation is not equally important in item 3.

The interview data does not indicate that these flaws actually have a negative effect on validity. Although some interviewed students definitely struggled with the complicated phrasing, none of these specific flaws were observed to cause interpretative problems. Consequently, the more complicated phrasing probably "only" results in students spending more time on these items compared to the original version. Nevertheless, as time is a limited resource, the phrasing should be reviewed.

11.3.2.2 *Motivation category 2 - Terms and notations*

An example for motivation category (2) are the "Representation" items. The term "partial free body diagram" is paraphrased to avoid confu-

sion with the concept of a full free-body diagram. The partial nature of the diagrams is mentioned separately as a note. This aspect was also addressed frequently by the interviewed experts, who see the partial representation as very problematic in the process of learning to properly use free-body diagrams.

See for example quotes on p. 134.

Another example is found in various items on the original CATS which involve the term "load(s)". The literal translation "Last(en)" was avoided, likely because it would be associated with external loads, only. Instead, the term "forces (and moments)" was used, again resulting in more text.

11.3.2.3 *Motivation category 3 - Conventions*

One of the examples for considering conventions is at the same time the only occasion where a sentence was omitted rather than added. In the Newton's Third Law items, the annotation that "forces act in the senses shown" was regarded as unnecessary by the translators as it is expected that German students generally assume arrows to indicate a specific direction of a force. Instead, arbitrary forces are unusual. Therefore, the translated version generally includes an additional note if arrows are arbitrary.

The FBD items show signs of all three motivation categories. FBD items 2 and 15 contain cords which are named in the figure in both versions. Only in the translated version, this name is referred to in the descriptive text for additional clarity and precision. Furthermore, weight and tension forces were renamed to match the local conventions (G instead of *W* for weight forces and S instead of T for tension forces), and an explanation of this nomenclature for the force types was added.

11.3.2.4 *Item 20*

Apart from the mentioned deviations which occurred repeatedly, various other changes were made in single occurrences, which will not be discussed in detail, except for item 20, which was investigated more closely. Item 20 features a subtle difference in translation, which does not affect the validity of the CATS in its current version (German or English), but it may have serious consequences for variants of item 20. The English version speaks of "All magnitudes are greater than zero", the translated version only says "not equal to zero". Together with the statement that "the forces and couple act in the directions and senses shown", both versions contain the same information, with a redundancy in the original. While removing this redundancy in the translated version was probably seen as an improvement, this small change can cause greater problems (in slightly modified variants of the item) than expected. The item expects students to conclude from non-positive magnitudes that equilibrium is impossible. Written work

from students on variants of this item shows that they often struggle to come to this conclusion. One variant required the magnitude of the couple to be negative, making equilibrium impossible under the constraints of non-negative magnitudes (see Chapter 12). The passage in the item stem "not equal to zero" is more often marked by the students than the passage "the forces and couple act in the directions and senses shown", indicating that the constraint of strictly positive magnitudes may be overlooked. In the CATS variant of item 20, undesired consequences of this rephrasing are *not* to be expected because equilibrium could only be established by illegally setting magnitudes to zero, i. e. negative magnitudes are not required.

11.3.2.5 *Summary of the translation analysis*

The translation analysis revealed some deviations from the original. The deviations were in part influenced by the instructors who had to give permission to administer the test in their courses. Their advice was valuable in terms of learning goals, conventions, and notations, but the tendency to phrase test questions (overly) precisely introduced a higher level of complexity. This tendency is probably culturally influenced, driven by the concern of legal consequences in case of ambiguously phrased exams.

Nearly all of the adaptations were implemented at the expense of not only creating a greater number of sentences but also more complicated phrasing. The construct "reading comprehension ability" becomes increasingly relevant with more complicated phrases, a construct which is not the desired one to measure. In combination with a tighter time limit, this can have a negative impact on the observed performance and thus on validity. Yet, the concerns raised here only partially align with results from the interviews, which will be presented in the following section.

11.3.3 *Student interviews (post-instruction)*

When making claims about the interpretability of the items, the target population should be consulted. The student interviews were conducted for this purpose, but with two foci.

The first focus of the post-instruction interviews is to test the hypothesis that the identified issues in the pre-instruction interviews can be purely attributed to the pre-instruction level of the students, i. e., the same problems would *not* be observed at post-instruction level. Thus, the item selection is based on the items in the pre-instruction interviews. Items which were possibly problematic in the pre-instruction interviews, i. e. items 1, 2, 4, 7, 9, 13, 20, and 25, were also examined in the post-instruction interviews, except for item 9. The reason for probing this item in the pre-instruction interviews was to find out more about the problems with item 2 and the term

"freigeschnitten". As will be shown, item 2 was not problematic in the post-instruction interviews, thus item 9 was omitted. Items which were never problematic and were shown in at least three pre-instruction interviews were replaced, i. e. items 10 and 19 (item 21 was kept to keep at least one Pin-in-slot item). Items which were never problematic but tested in less than three pre-instruction interviews, i. e. items 8, 14, and 21, were kept.

The second focus of the student interviews is to test various hypotheses generated from the expert interview results. A few interviews are of course not sufficient to come to reliable conclusions on each matter but they provide a first insight. Items 5, 16, and 17 were added to the selection with this focus. In item 5, the point of interest was misinterpreted by experts. For item 16, a result from the distractor analysis in Section 11.2.2 is the hypothesis that the frame elements in the Newton's Third Law items are misinterpreted as two-force members such as rods in trusses. Finally, item 17 was added because item performance is poor, and the graphical element of the dashed lines indicating the right angle was missed in the item selection for the pre-instruction interviews.

The experts also mentioned issues on items that were already in the selection. Apart from the problems identified in the pre-instruction interviews, the effect of the missing weight of the block in item 1 on the students' thoughts are examined. Likewise, the students' reactions to the support of block 4 in item 2 as well as the arm in item 8 and on the special (i. e. horizontal) direction of the forces in the correct response of item 4 are analyzed.

Table 5 shows an overview which items were presented in each post-instruction student interview and an evaluation of how well the item was interpreted by the student. Most items were always correctly interpreted but the analysis shows more or less frequent problems in items 4, 8, 16, 21, and 25. The results confirmed some but not all of the interpretative problems anticipated by the experts.

11.3.3.1 *Instructions*

The experts noted that stating the instructions only once at the beginning may not be sufficient. Usually, this was not a problem for the interviewed students. Only in a few cases in the later interviews (e. g. student S12 item 6, S13 item 17, S14 item 16, S15 item 6), the interviewer decided to hint at the instructions, because they were not consulted or recalled by the students themselves.

Frequently, the meaning of the word "pin" ("Stift" in German, also meaning "pen") in the instructions was unclear before seeing the items.

> *"Is 'Stift' now a pin that can be inserted through a joint, or a pen for writing? It's the pin you put through a joint, isn't it?"*

Table 5: Interpretation of the selected items in the student interviews: correct interpretation or minor misinterpretation (●), initial misinterpretation but corrected (◑), major misinterpretation (○), inconclusive (?). (Items not selected for any interview are not listed.)

Item	Student Interview															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		●		●							●					
2	●		●			●		●								
4	○	○					●?		●							
5			●				●				●					
6												●	○		●	
7			●	●			●							●	●	
8					○					◑	○			●		
11													●	●		●
13			●					●			●					
14																●
16			●					○	●					◑	●	
17			●			●						●	○		●	
20		◑	●		●		●		●	●		●	●		●	
21						○		?				●				●
23					●				●			●				
25	◑			○					●?	○					●	

> *"Ist mit 'Stift' jetzt ein Stift gemeint, den man durch ein Gelenk durch stecken kann, oder ein Stift zum Schreiben? Es ist der Stift, den man durch ein Gelenk steckt, oder?"*
>
> *S2 #00:23:09#*

The confusion got resolved once the students encountered a pin in the item context. This was anticipated by some experts, e. g. E3:

> *"'Pins', for example, we don't use that expression. That means that, at first, the students wouldn't know exactly what is meant by that. Of course, I would think that they would relatively quickly resolve what is meant by that and then draw conclusions about how the forces must act where such pins are inserted."*
>
> *"'Stifte' zum Beispiel, den Begriff verwenden wir gar nicht. Das heißt, die Studierenden wüssten erstmal nicht so genau, was damit gemeint ist. Natürlich würde ich schon denken, dass sie sich das relativ schnell selber klar machen, was damit gemeint ist, und dann auch wieder Rückschlüsse ziehen, wie denn dort die Kräfte auftreten müssen wo solche Stifte eingesetzt sind."*
>
> *E3 #00:05:26-3#*

One exception was S14 on item 16 (Newton's Third Law), where the resolution was only found after the interviewer gave an additional hint at the instructions. This interview is described subsequently in more detail as it involves three interesting aspects: (1) the misinterpretation of "pins", (2) the misinterpretation of the members as rods, and (3) the possible ineffectiveness of the Newton's Third Law items to test understanding of Newton's Third Law.

11.3.3.2 *Item 16*

Item 16 was selected to check whether students tend to misinterpret the structure to be a truss made of two-force members. This kind of misinterpretation was observed once. S14 interpreted the term "pins" as referring to the connected members (S14 #00:08:17#) which he assumed to behave like rods, i. e. forces can only act along the members. The student stated clearly that the forces need to be opposite, and was confused when he found no response option allowing for both, forces along the members and opposite to each other (which is geometrically impossible).

> *"So, now I assume that such a pin acts like a rod, so that it also can take only forces in longitudinal direction and not in transversal direction like a beam, for example. And therefore, the only option would be (a), [...], but still, in case of (a), I could not explain, if it [the pin] was a rod, why the forces are not opposite [...]."*

> *"Also ich nehme jetzt an, dass so ein Stift so in der Art wie ein Stab wirkt, also dass er auch nur Kräfte in die Längsrichtung aufnehmen kann und nicht jetzt in die Querrichtung wie jetzt ein Balken zum Beispiel. Und demnach wäre ja nur die Möglichkeit (a), [...], aber dennoch könnte ich mir bei (a) dann nicht erklären, wenn es [der Stift] ein Stab wäre, dass die Kräfte ja trotzdem nicht entgegengesetzt sind [...]."*
>
> *S14 #00:07:15#*

Only when the interviewer hinted at the instructions, the confusion about the pins got resolved. The question remained of which loads the depicted members could support. He expressed doubt that his interpretation of the members as rods is correct (S14 #00:09:27#). When he was forced to commit to a response, he chose the correct one (S14 #00:09:59#). The condition that the forces must be opposite weighs more strongly for him than the (mis)interpretation of the members as rods.

Still, it is unclear whether S14 distinguishes between a Newton's Third Law force pair (where the forces act on different bodies and are always opposite and equal by law of nature), and two forces acting in opposite directions on the same body, (which may or may not be equal in size such that the body may or may not be in equilibrium, depending on external loads). He never explicitly mentioned Newton's Third Law but, earlier in the interview, he names two reasons for opposing forces: conditions for "fulfilling equilibrium" and conditions for "fulfilling support reactions". The latter, a reaction by the support, hints at Newton's Third Law:

> *"Usually, [the forces] are in opposite directions in order for equilibrium to be fulfilled, or in order for this support reaction to be fulfilled."*
>
> *"Normalerweise sind die [Kräfte] in entgegengesetzter Richtung, damit das Gleichgewicht erfüllt ist, oder damit diese Lagerreaktion erfüllt ist."*
>
> *S14 #00:06:50#*

Despite knowing about these two reasons for opposing forces, S14 never justified his choice for the correct option with one or the other. In case he does not distinguish between those cases when responding to the item, the response could be a false positive. The Newton 3 item(s) would not actually measure understanding of Newton's Third Law. Unfortunately, the data do not allow further insight, but this question should be investigated in the future.

S8 was observed to produce a false positive response (though tentatively), reasoning that the direction of the forces is linked to the

direction of the slot (which is entirely irrelevant in this case). This could be easily avoided by changing the direction of the forces in the correct response to not align with the slot.

11.3.3.3 *Arbitrary forces*

Other issues frequently occurred in the interpretation of "arbitrary forces" (e. g. items 4 and 16), even though the arbitrary nature of the forces was emphasized even more in the translated version. If the correct response was not easily found, students often attempted to identify it from the direction of the arbitrary forces.

> *[Chooses 4(a)] "I looked at it and just saw, ok, here a force acts downwards and there one to the left and in the picture (a) there's one that acts upwards and to the right. And with this, I would have my equilibrium conditions, which could roughly balance it, [...]."*
>
> *[Wählt 4(a)] "Ich hab mir das angeschaut und hab halt gesehen, ok hier wirkt ne Kraft nach unten und da wirkt eine nach links und im Bild (a) hab ich eine, die wirkt nach oben und nach rechts. Und damit hätte ich in meine Gleichgewichtsbedingungen, die das ungefähr ausgleichen könnten, [...]."*
>
> *S2 #00:24:30-2#*

Some students interpreted the attribute "arbitrary" to apply only to where the force acts or to its magnitude, but not to its direction (e. g. S2 #00:26:08-9#, S1 #00:09:35-7#).

11.3.3.4 *Two-dimensional force vectors and "reactions"*

Item 25 asks for possible "reactions", shown as arrows, at a frictionless pin support. The question was sometimes misinterpreted to ask for reactions in terms of possible movements, although an additional hint is given in the item question that forces and couples are represented by arrows. S10 correctly interprets the straight arrow as a force and labels it F. He also correctly interprets the support as a hinge, using gestures to indicate the only degree of freedom, and draws two separate free-body diagrams for the lever and the support, which are also correct (they both include a vertical and a horizontal force component, and the magnitude and direction of the forces are consistent with action = reaction). He then adds F as an *additional* external load to the free-body diagram of the lever and says that F can be calculated from the equilibrium conditions. He even draws a right-angled triangle to illustrate that the two components of the support reaction must be equal to the horizontal and vertical component of F. He thus does *not* interpret F to be the resultant of the components of the pin support reactions, but as an additional force, which can be compensated by the pin support such that no reaction (i. e. movement) occurs.

→ *Translation Analysis in Section 11.3.2.*

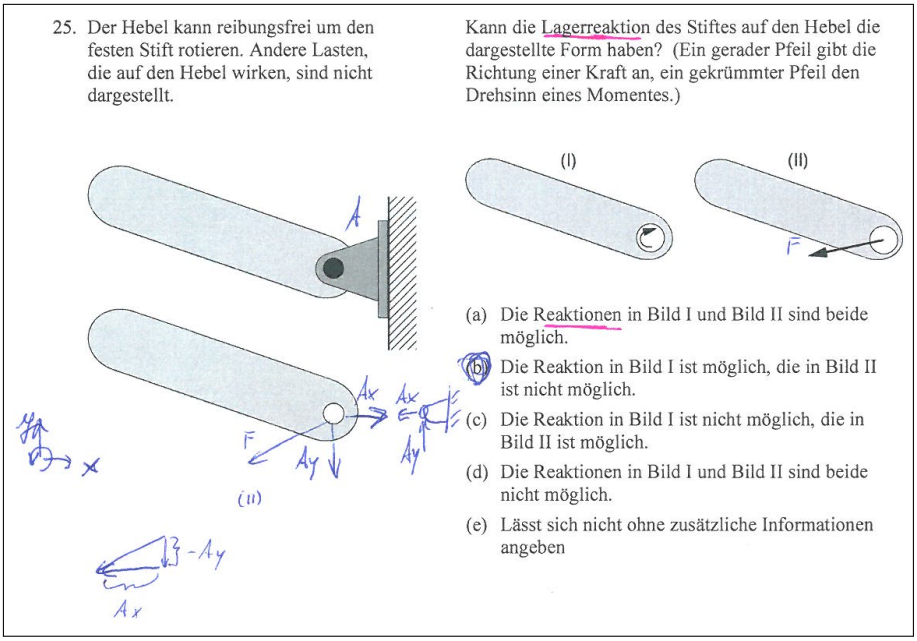


Figure 21: Notes made by S10 during the interview on item 25. (Note: The underlines in the text were drawn by the interviewer after the interview.)

S10: "So I had chosen (b) [...] because of (reads) *'a straight arrow indicates the direction of a force, a curved arrow the direction of rotation.'* (incorrectly quoted) *If we say, or if we want to look at which of the movements would be possible, then this would be possible* (points at (I)), *but this would be* (points at (II)), *well, the system wouldn't move if this force would act."*

I: "[...] What is meant by 'reactions' here in the possible answers? Or how do you interpret that?"

S10: "(reads) *'The reactions in picture (I) and picture (II) are both possible'. [...] I interpret this as the possibility of movement or the possibility that something happens in the given case [...]."*

I: "But related to movement, in a way."

S10: "Yes, related to movement."

> *S10: "Also (b) hatte ich gewählt, [...] wegen* (liest) *"ein gerader Pfeil gibt die Richtung einer Kraft an, ein gekrümmter Pfeil die des Drehsinns."* (inkorrekt zitiert) *Wenn wir hier dann sagen, oder wenn wir betrachten wollen, welche der Bewegungen möglich wäre, dann wäre die hier möglich* (zeigt auf (I)), *aber die hier* (zeigt auf (II)) *wär halt, also das System würde sich so nicht bewegen, wenn da diese Kraft wirken würde."*
>
> *I: "[...] Was ist mit "Reaktionen" hier in den Antwortmöglichkeiten gemeint? Oder als was interpretierst du das?"*
>
> *S10: "*(liest) *'Die Reaktionen in Bild (I) und Bild (II) sind beide möglich.' [...] Ich interpretier das als Bewegungsmöglichkeit oder die Möglichkeit, dass etwas passiert in dem vorgegebenen Fall [...]."*
>
> *I: "Aber bezogen auf Bewegung, sozusagen?"*
>
> *S10: "Ja, auf Bewegung bezogen."*
>
> *S10 #00:10:03-4# ff.*

Using the term "support reactions" (= "Lagerreaktionen") as in the stem instead might avoid confusion.

A similar problem of confusing the meaning of arrows as indicating motion occurred in Limit of Friction item 6. Even though the text states that forces are depicted and the arrows are even labeled with the unit N for Newton, one student interpreted the arrows as velocities (S13 #00:09:27#).

Similarly to the problem with arbitrary forces, two students (S4 and S9) struggled to interpret the force in item 25 as one of many possible forces instead of a specific (albeit arbitrarily chosen) solution to the question. They tried to find an explanation for the depicted directions of both, the moment and the force. When none was found, they concluded that the pin connection must resemble a movable bearing because the support can only exert a force in one direction.

> *"Yes well, now the question is... why the moment is drawn only in one direction, because if it can rotate frictionless in this point around the fixed pin, it can rotate in the one direction as well as in the other. Therefore I wonder why it is only the one moment in (I). Yes, well, the direction of a force, well, can be a possible one.* (unintelligible) *Well ... I would interpret it as a movable bearing. Then one would actually only have one effective force. That is, if anywhere (II), but ... why exactly this [force] is not quite obvious to me."*

> *"Ja gut, jetzt ist ja die Frage... warum das Moment nur in die eine Richtung eingezeichnet ist, weil wenn es sich hier ja in dem Punkt reibungsfrei um den festen Stift rotieren kann, kann es ja auch sowohl in die eine Richtung als in die andere Richtung rotieren. Deshalb frage ich mich, warum es in (I) nur das eine Moment ist. Genau, also die Richtung einer Kraft, ja gut, kann eine mögliche sein.* (unverständlich) *Ja ... Würde ich so als Loslager theoretisch einschätzen. Dann hätte man ja eigentlich nur eine wirkende Kraft. Das heißt, wenn dann in Bild (II), aber ... warum genau diese [Kraft] ist mir nicht ganz ersichtlich."*
>
> *S9 #00:18:47-7# f.*

The students did not interpret the force as the vectorial sum of any support reactions in horizontal and vertical direction. This aspect was also anticipated by some experts and it is (at least in part) driven by instruction culture: The students reported that they are explicitly taught to always draw perpendicular force components and not the resultants (in terms of the vectorial sum), a fact that was independently mentioned by the instructor of the interviewed students as one of the interviewed experts.

> *"So I'm not sure about this arrow here [in picture (II)], because I am used to it [...] from the exercise tasks, [...] that I draw [a force], well, [...] in x-direction or y-direction and here it [the arrow] somehow goes down slanting [...]."*
>
> *"Also mit diesem Pfeil hier [in Bild (II)] kann ich irgendwie nicht so viel anfangen, weil ich bin es [...] von den Aufgaben immer gewohnt, [...] dass ich [eine Kraft] halt [...] in x-Richtung oder y-Richtung einzeichne und hier geht er [der Pfeil] irgendwie schräg nach unten [...]."*
>
> *S1 #00:28:31-5#*

Drawing support reaction forces as components makes sense in case the true direction of the force is yet unknown. Being confronted with a type of conceptual question that she was not "used to", revealed a severe lack of functional understanding of vectors: When asked subsequently whether it was possible to represent the angled force by two forces, one in horizontal and one in vertical direction, S1 stated that it was possible, but she could not do it.

11.3.3.5 *Item 8*

The question on item 8 seems to be difficult to understand. It asks whether a force acting between two shown parts can possibly act in the direction "perpendicular" or "parallel" to one of those parts. Only

one of four students correctly interpreted the item immediately. One student (S11) initially misunderstood the question, thinking that it is asks whether the drawn situations are equivalent.

> *"The question now is [...] whether one can also represent this with figure (II). So figure (I) actually shows exactly [...] a part of the image in the [stem of the] question and [...] figure (II) is a different figure [situation]. The question is whether this is possible. (I) is clearly possible."*
>
> *"Die Frage jetzt ist [...], ob man die Darstellung auch mit Abbildung (II) zeichnen kann. Also Abbildung (I) zeichnet eigentlich genau [...] eine Teilabbildung von der Abbildung in der Aufgabe und [...] bei der Abbildung (II) handelt es sich um eine andere Abbildung. Die Frage ist, ob das möglich ist. (I) ist schon klar."*
>
> *S11 #00:12:41-9#*

He sees situation (I) to be equivalent to the item setting and says that the force of the arm on the lever acts along the arm.[6] After a while, he notices his mistake and correctly interprets what the item asks for. Still, the entire setting is not quite clear. S11 did not specifically address the support of the arm, but the type of support seems to be clear judging by his other arguments. He stated that the indicated moment originates from the arm opposing the lever's weight force.

> *"So there's a weight force here [at the rocker]. If there was no moment here [at the arm], then this weight force would push the arm [down] so that [it] rotates on [the] axis. This moment here* (unintelligible) *is [the] moment generated by the weight force."*
>
> *"Hier [beim Hebel] gibt's also eine Gewichtskraft. Wenn kein Moment hier [beim Schwenkarm] gibt, dann würde diese Gewichtskraft den Schwenkarm so [nach unten] drücken, dass [er] auf [der] Achse rotiert. Dieses Moment hier* (unverständlich) *also [das] erzeugende [erzeugte?] Moment von Gewichtskraft."*
>
> *S11 #00:22:57-3#*

This alternative interpretation alone would not lead to serious consequences (because for the concept of interest it does not matter whether the contact force is ultimately caused by the spring or by gravity). He selects the correct response, but for the wrong reason, arguing that the force must be vertical because it opposes the weight force of the lever. In this case, the item produced a false positive result,

6 It should be noted that S11 was not a native speaker. His German was not free of errors but he was speaking fluently so that it is assumed that this mistake is not necessarily caused by language problems.

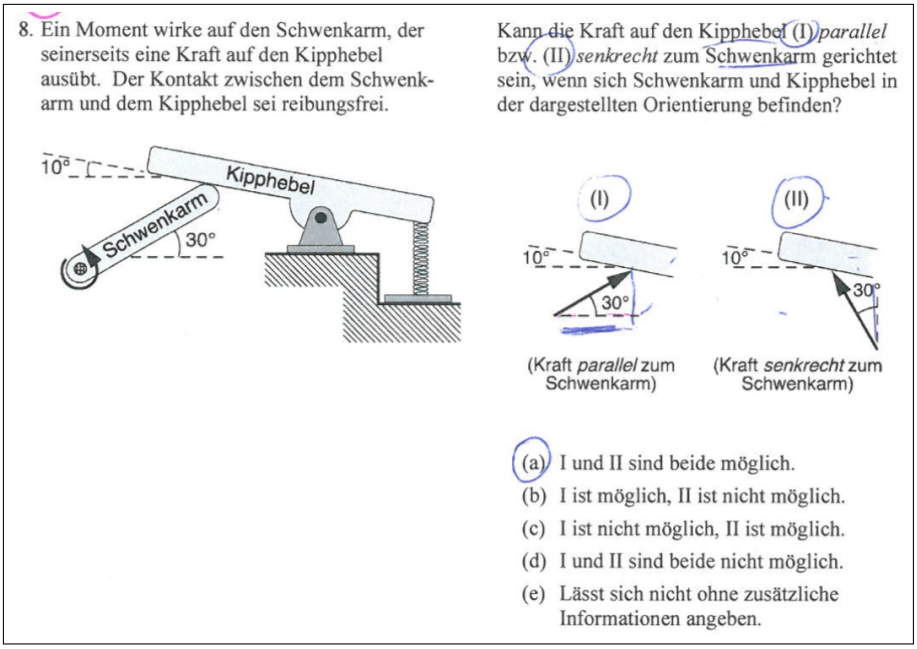


Figure 22: Notes made by S5 during the interview on item 8. S5 interpreted (I) to show the lever ("Kipphebel") and (II) to show the arm ("Schwenkarm"). He also marked "parallel" as horizontal in (I) and "perpendicular" ("senkrecht") as vertical in (II).

caused not by misinterpretation but by lack of conceptual understanding of the possible directions of forces acting at surfaces of negligible friction. The idea that the force must be directed vertically to oppose the force by the lever is not reflected in the distractors.

S5 initially interpreted (I) to show the lever ("Kipphebel") and (II) to show the arm ("Schwenkarm"). To avoid this misunderstanding, the reference to the images (I)/(II) should be placed *after* the words they refer to[7]. S5 also seems to think that the term "dargestellte Orientierung" (given orientation) refers to the figures below instead of the figure on the left. He also had problems to interpret from both, the text and the drawings, which directions are specified by "perpendicular" and "parallel" in this case. When asked to indicate these directions, he marked "parallel" as horizontal in (I) and "perpendicular" ("senkrecht") as vertical in (II). For him, the item seems to ask whether it is possible to calculate the horizontal and vertical force components or change the direction of the forces such that it acts horizontally (in case (I)) or vertically (in case (II)).

> *"Well, I'd answer (a), that both are possible. Well, I imagine that I can, because of having the angle data, that I can change the force, let's say with trigonometry, so that it is parallel or perpendicular."*

7 i. e. "Kann die Kraft auf den Kipphebel *parallel* (I) bzw. *senkrecht* (II)..."

> *"Also ich würde halt (a) antworten, dass beides geht. Also, ich stell mir das so vor, dass ich das dann halt, dadurch, dass ich die Angaben vom Winkel habe, dass ich die Kraft, sagen wir mit der Trigonometrie, so ändern kann, dass die parallel oder senkrecht ist."*
>
> *S5 #00:10:05-2# ff.*

It was not clear from the interview whether he really believes that he can physically change a force by trigonometry or just calculate the force component in a certain direction, but he used this expression more than once.

S10 also misinterpreted the directions as parallel/perpendicular to the lever, instead of the arm. The directions may be difficult to interpret because the arm as the direction-giving object and the force in question are not drawn in the same picture. To eliminate the problematic terms, the drawings could be modified. Instead of showing only one part of the setup and the force acting on it, the interacting bodies could both be shown separately with the force pair acting between them.

See expert interviews (→ p. 134)

As anticipated by the experts, one student was briefly confused by the representation of the support of the arm. S10 hesitated for a moment but then the situation was clear to him.

Students often assumed the setting to be in motion. This may again be induced by a confusion of the arrow indicating a moment or a motion. Alternatively, the students may not see that there can be one without the other in this situation, which would be a common misconception that is also addressed by the FCI (Hestenes et al., 1992).

11.3.3.6 *Item 20*

See also Chapter 12.

The most problematic item in the quantitative analyses discussed below (Section 11.3.4 - Section 11.3.6), item 20, was shown to nine students, and was most often *correctly* interpreted. Only two students misinterpreted the term "Drehsinn" (sense of direction of the couple) as the resultant moment of the acting forces or the resulting movement of the body, a misinterpretation which would not affect their response. One student (S2) spotted that, technically, this item is incompatible with the instructions. He referred to the statement *"Alle Körper befinden sich im statischen Gleichgewicht"* (all bodies are in static equilibrium) and concluded that, consequently, both bodies must be in equilibrium. It is assumed that this reasoning is not applied frequently, but it would be interesting to know how many students would actually base their response on the same argument to see if the question or instructions must be changed.

11.3.3.7 *Other items*

Item 1: The missing weight force was never noticed by the students.

Item 2: Two of four students (S3 and S6) misinterpreted the weight force as not explicitly given and thus negligible. Due to the given response options, this misinterpretation was quickly resolved. The possibility of block 4 falling down was never mentioned.

Item 5: The point of interest, which was repeatedly misinterpreted by experts (→ p. 134), was clear in all three student interviews on this item.

Item 6: One out of three students misinterpreted the force in the response options as the sum of all forces. The notation of the blocks' weight forces was never a problem.

Item 7: One out of five students started to look for a possible application (as observed in the pre-instruction interviews, → p. 77). The same student suddenly assumed the item to be multiple-select.

Item 21: S6 struggled with the complicated phrasing and could not identify the relevant force. S8 may have thought that the item asks for the internal moment in the slotted element, but the data is inconclusive.

11.3.3.8 *Summary of the student interviews*

Most of the problems that pre-instruction students encountered did not occur in the post-instruction interviews. For instance, the meaning of technical terms such as "torque" or "free-body diagram" are generally clear by then. These problems show that the CATS is not suitable as pre-test as they can be attributed to the pre-instruction status of the students. Two problems which still occurred in the post-instruction interviews is the confusion of arrows as indicating movement instead of forces or torques, and the concept of arbitrary forces. This problem is hence independent of whether or not students experienced formal instruction. In addition, some of the problems anticipated by the experts, such as failure to interpret arrows as complete force vectors, could be confirmed in the student interviews. These aspects reduce the precision of the measurement, as the understanding of none of these concepts (arbitrary forces, arrows representing forces) is explicitly intended to be measured by the CATS. Especially the concept of arbitrary forces is merely used as a tool to focus the items on the concept of the respective category and discourage quantitative approaches, it is not the intention of the CATS to measure the understanding of this concept. If students miss an item because they do not understand arbitrary forces, it suggests that the CATS measures other constructs than intended.

Recognizing arrows as indicating forces and the concept of two- (or three-)dimensional force vectors resulting from components, however, are both fundamental concepts to understanding statics, as can be concluded from the axioms and the importance the experts attributed to the concept of vectors. If students miss items because they did not understand either of these concepts, this is an indication that they have a poor understanding of statics. Student's problems with interpreting arrows as forces and resulting force vectors are therefore supporting evidence that the *overall construct* to be tested is indeed conceptual understanding of statics. However, as these concepts appear in items of various CATS concept categories, student issues with these concepts limit the instrument's power to accurately measure student understanding at the sub-scale level. Furthermore, these issues may limit the power of the CATS as a diagnostic tool for misconceptions.

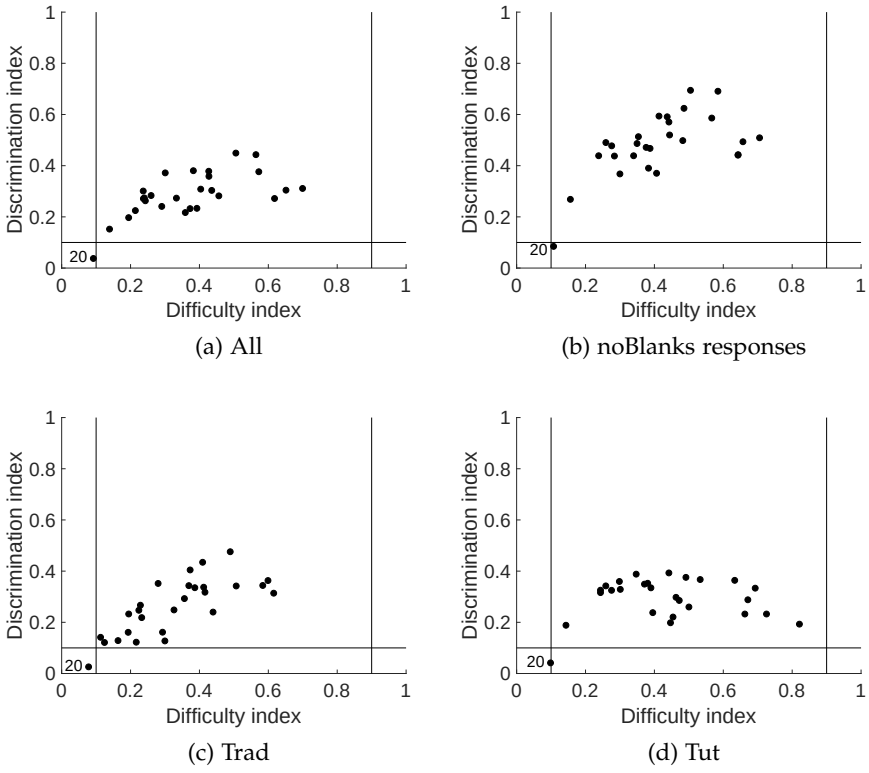
Most of the reported results elaborate on student *issues* with the items, because misinterpretations are of course more interesting and they require more explanation than correct interpretations. This imbalance may leave the reader with the impression that the interpretability of the CATS is weaker than it actually is. It must hence be emphasized that, overall, most of the selected items were correctly interpreted by the students. Especially the most problematic item in the quantitative analysis, item 20, was most often correctly interpreted. The one misinterpretation that did occur was harmless in that it did not affect the students' response. The remaining problems to find the correct response were due to lack of conceptual understanding of statics and are therefore a positive indicator for validity.

11.3.4 *Classical Test Theory analysis*

The previous investigations addressed the aspect of interpretability of the items and were entirely based on qualitative data. The quantitative results of the validation analysis presented in this and the following sections are based on the framework proposed by Jorion et al. (2015), which involves inspection of item statistics and reliability in terms of a CTT analysis (presented in this section), as well as IRT and factor analysis (presented in the following sections).

11.3.4.1 *Item statistics*

Figure 23 shows the distribution of difficulty and discrimination indices for the different data sets. According to the scheme proposed by Jorion et al. (2015) (see Table 9), acceptable values for item difficulties lie between 0.1 and 0.9, and discrimination values are merely required to be positive.



(a) All (b) noBlanks responses

(c) Trad (d) Tut

Figure 23: Difficulty and discrimination indices in CTT analysis

STANDARD DATA SET In the standard data set (Figure 23a), all items have acceptable difficulty and discrimination with the exception of item 20, which is too difficult and shows signs of low discrimination. The difficulties of the other items spread out across a large part of the range of acceptable values, indicating a good composition of the test in terms of easy and difficult items. The overall discrimination power of the test is concentrated in the lower half of the acceptable range.

NO BLANK RESPONSES Excluding student data with blank responses largely increases all items' discrimination indices because an item that is not responded to cannot discriminate at all between individuals. When excluding student data sets with blank responses, all items are in the acceptable ranges, with the exception of item 20 which remains low in discrimination (Figure 23b).

The difficulty indices are less affected by blank responses. (The maximum difference to the standard data set in terms of difficulty is 0.02.) For the early items, the difficulty index or item score tends to be lower in the noBlanks data set compared to the standard data set (not visualized). This finding contradicts the hypothesis that the "fast working" students in the filtered data set are in general stronger performers. For the later items, the difficulty index or item score tends

Data set	Cronbach's α	$\alpha_{-i} > \alpha$
All	0.80	[20]
noBlanks	0.81	[20]
Trad	0.76	[20, 27]
Tut	0.79	[20]

Table 6: Cronbach's α for the different subsamples and reference values from literature

to be higher in the noBlanks data set compared to the standard data. The reason for this is that blank responses mainly affect the late items which artificially reduces the item score if blank responses are graded as incorrect.

SEPARATE PEDAGOGIES When investigating the effect of instructional method, item 20 is also in the problematic area in both cases. For the Trad case, several other items with comparably low difficulty index are close to critical discrimination (Figure 23c). Unlike in the Tut case, no item difficulty index is substantially higher than 0.6 (Figure 23d). The CATS seems to perform substantially better in case of interactive instruction[8].

11.3.4.2 *Reliability*

Reliability of a scale was introduced in Section 9.3.1.1 as a central concept of CTT that is often assessed with Cronbach's α, a statistic which is based on the concept of internal consistency. It assumes that all items measure one single construct. This demand for a one-dimensional scale competes against the demand for a clear structure of different concepts. On a test with high internal consistency, the responses to all items correlate positively with one another. A perfect correlation of 1.0 would indicate that any single item from the CATS would be sufficient to test the construct. With supposedly nine concepts to test for and a reasonable result for the proposed concept structure (see factor analysis below), a value close to 1.0 is not expected. Values for α in an acceptable to good range from 0.76 to 0.81 were found (see Table 6).

As α is sensitive towards the total number of items N, it is inspected whether the internal consistency improves or stays the same if an item i is removed ($\alpha_{-i} \geqslant \alpha$). In this case, item i probably measures a different construct or is badly phrased. $\alpha_{-i} \geqslant \alpha$ was found for item 20 in all data sets, and additionally for item 27 in the Trad data set (Table 6).

8 Primarily, it is the students who perform better, but that is not the question that is relevant here

According to the population dependence, which was mentioned in Section 9.3.1.4 as one of the shortcomings of CTT, the reliability for the CATS over the joint population of all cohorts is expected to be larger than the ones over only the Trad or only the Tut cohorts, a result which is indeed observed here. This is because the variance of scores is larger for the joint population as the two subgroups differ in performance.

CTT summary:

- Item 20 is problematic in all cases.
- Removing blank responses increases discrimination.
- The CATS seems to be better suited to discriminate between individual students in case of Tutorial instruction.
- Reliability is acceptable.

11.3.5 *Item Response Theory analysis*

In the following sections, the results of the revalidation in terms of IRT are discussed.

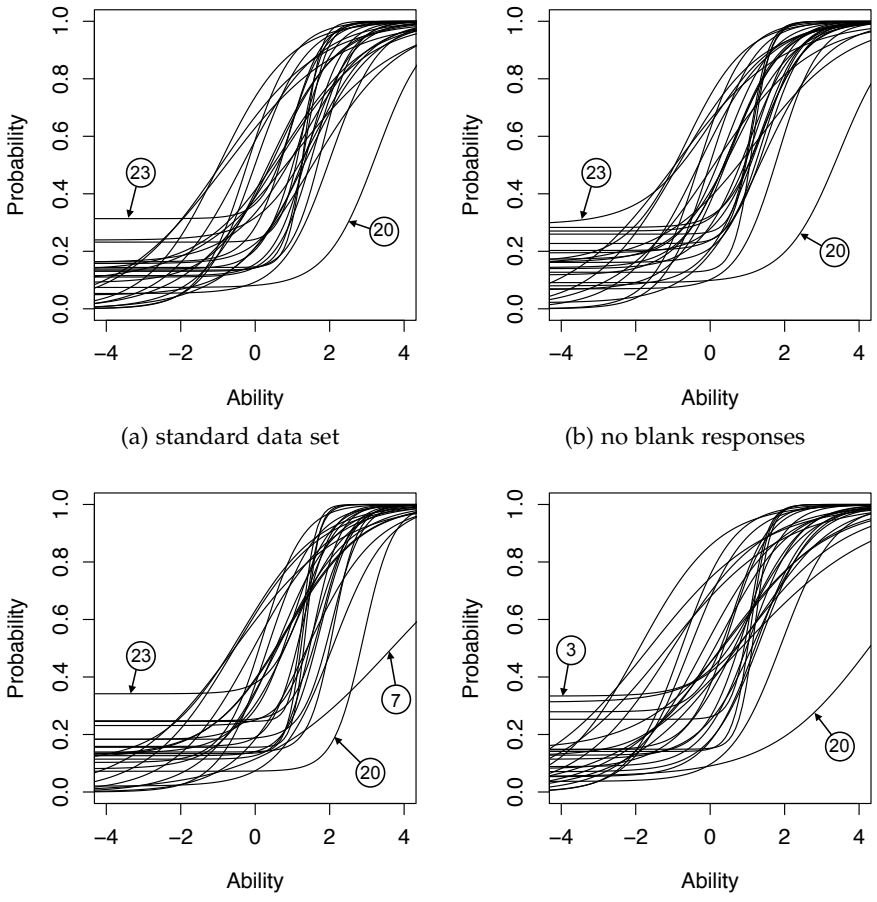
11.3.5.1 *Model selection*

The model selection should be informed by theory and statistics. By theory, a 3PL-model may be chosen in order to include a guessing-parameter because it can be assumed that the MCQ-format in combination with a very ambitious time limit fosters guessing behavior, especially for late items. On the other hand, a 2PL-model may also be justified as the students do not benefit from guessing because the test is not graded and therefore may choose not to guess. However, the Akaike information criterion (AIC), which was consulted also by Jorion et al. (2015) to determine the model with the best fit (while penalizing model complexity), suggests that the 3PL-model fits best in this case.

11.3.5.2 *3PL-model evaluation*

Figure 24a shows the ICCs of all items for the standard data set. Item 20 is noticeably more difficult than the rest of the items. This result is in accordance with the CTT analysis.

Some curves do not have a strong asymptote on the plotted ability range. These could well be described by a 2PL-model. Thus only some items are affected by guessing. Especially item 23 stands out with about a one to three chance of guessing the correct response.



(a) standard data set (b) no blank responses

(c) traditional instruction (d) interactive instruction

Figure 24: 3PL-model item characteristic curves. Item numbers in circles indicate noticeable items in the IRT analysis.

Looking only at data without blank responses (Figure 24b) results in more items with higher guessing probabilities. It seems that this subgroup is more successful at guessing. This might be an effect of the sample characteristics and their test taking strategy, as aiming to answer all items in time encourages guessing behavior.

Analyzing the data with pedagogies separated reveals the following differences: Item 23 is the only item with high guessing probability in the Trad sample. In the Tut sample, there are several items with a higher than expected guessing probability, led on by item 3. The most difficult and poorly discriminating items are item 7 on the Trad sample and item 20 on the Tut sample. Item 20 is also very difficult on the Trad sample but discriminates well. Unfortunately, such items are responsible for shifting the focus of the test too far to the higher abilities. This will be illustrated in the following paragraphs.

11.3.5.3 *Test information function*

The test information function for the standard data set is shown in Figure 25 together with a histogram of the estimated abilities. Because the focus of the revalidation is not limited to a certain ability range, the ideal test information function in the context of this work would be a constantly high value, but this is "hard to achieve" (Baker, 2001, p. 105). A test information function with a very moderate peak at $\theta = 0$ would be a more realistic goal.

Unfortunately, the CATS is most informative, i. e. measures most precisely, for higher abilities in the range between 1 and 2. The function exhibits very low values for abilities below 0. As guessing is considered, low values are expected towards the very low ability range, but the low precision in the moderately low ability range between -2 and 0.5 is undesirable, as many students' abilities lie in this range. The drop in precision towards the higher end is acceptable as few students exhibit very high abilities.

Figure 26 shows the test information functions for the different subsets. Including all students in the sample allows to measure more precisely over the largest part of the plotted ability range than for any of the subsets alone. This is reasonable as the sample is more diverse than any of the subsets. Considering only the noBlanks students leads to a less precise measurement than with all students, especially for abilities above 1.

For the Trad subset, the peak is comparably high as or the All sample, but it drops more quickly towards the lower abilities. The test information values of the Tut sample is much lower, not only in the peak, but also towards the higher abilities. Only in the medium to low range abilities, they are slightly higher than the Trad curve.

The difference in the peak positions indicates that the average item difficulty is greater[9] for the Trad than for the Tut students. In the Trad sample, the CATS' focus lies even further out on the high ability range and it measures more precisely in this range than in the Tut sample. Unfortunately, only very few students are found in this ability range.

CONCLUSION Applying the CATS in the given context allows for high precision measurements only for the high ability students. This is especially true for the Trad sample. For the majority of students, the results show a weakness in precision. Considering all abilities, the CATS may be slightly more suitable for the Tut than for the Trad students. This corresponds to the conclusion from the CTT analysis.

9 Attention: In IRT, "greater difficulty" means indeed "harder", unlike in CTT.

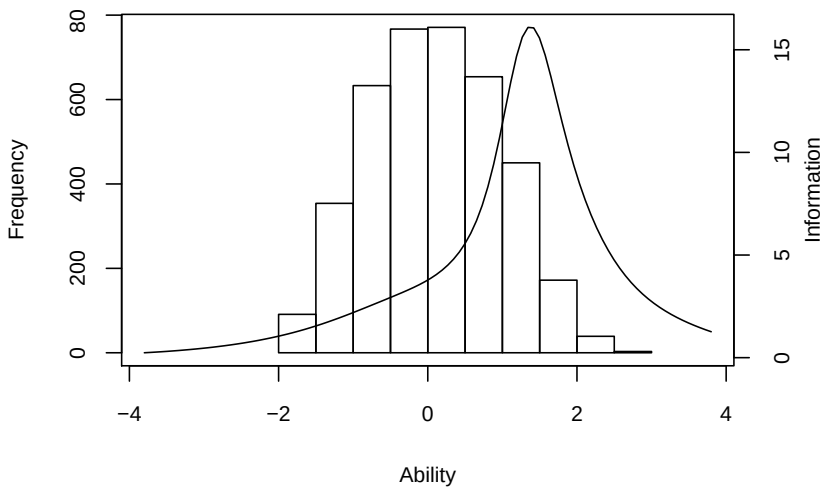


Figure 25: Test information function and histogram of student abilities for the standard data set

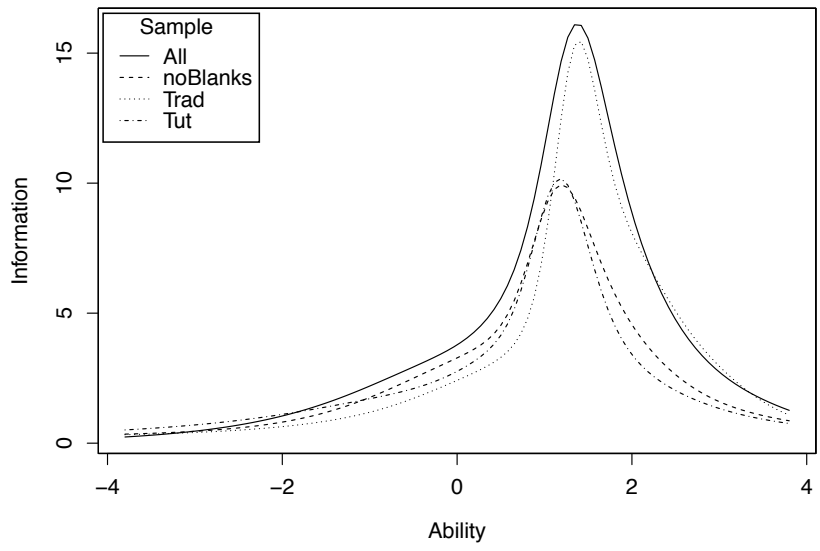


Figure 26: Test information functions for all inspected data sets

11.3.6 *Factor analysis*

A factor analysis is applied to investigate the structure of a test scale and check for subscales. As CTT analysis revealed the following items to be problematic. They will be omitted in this structural analysis:

Data set	Problematic items
All	20
noBlanks	20
Trad	20,27
Tut	20

The proposed structure of the CATS expects nine factors, a parallel analysis as introduced in Section 9.3.4 suggests eight factors only, indicating that not all proposed concepts can be found in the data.

Table 8 shows the factor loadings on the eight factors for the standard data set. Because many of the expected items grouped together, the factors were named according to the predefined concepts. Six factors could be identified to align with the proposed concepts. A concept is defined here as identified, if at least two concept-related items and no unrelated items load onto the same factor. The concepts *Frictionless contact*, *Representing loads* and *Equilibrium* could not be identified in case of the standard data set. (If nine factors are assumed, the result is the same.)

Although not on the same concept, items 23 (*Representing loads*), 25 (*Frictionless contact*), and 26 (*Equilibrium*) load onto the same factor. Item 25 is indeed closely related to the representation of loads at connections, because it is asking for what kind of reaction (force or couple) can occur at a frictionless pin support, which is a standard type of connection. A conceptual explanation why item 26 loads onto the same factor could not be found. There is no theoretical connection between these items except that these are all late items, so blank responses might be a part of the explanation. When removing data with blank responses item 26 does not factor with items 23 and 25 (or any other items, see Table 16).

Overall, removing data with blank responses results in a slightly better representation of the predicted structure. The *Frictionless contact* concept could additionally be identified. As in case of the standard data set, the concepts *Equilibrium* and *Representing loads* do not emerge from the data (see Table 16).

Investigating the pedagogies separately reveals quantitative and qualitative differences. Note that in the Trad case, more items were omitted. Nonetheless, the parallel analysis suggested eight factors. For Trad, six concepts could be identified like in the standard case but instead of *Static equivalence*, *Frictionless contact* was identified. For

Table 7: Identified concepts in the factor analysis

Concept	All	noBlanks	Trad	Tut
Drawing forces on separated bodies	✓	✓	✓	✓
Newton's Third Law	✓	✓	✓	✓
Static equivalence	✓	✓	-	✓
Roller joint	✓	✓	✓	✓
Pin-in-slot joint	✓	✓	✓	✓
Frictionless contact	-	✓	✓	✓
Representing loads at connections	-	-	-	-
Limits on friction force	✓	✓	✓	✓
Equilibrium	-	-	-	-

Tut, seven concepts were identified, including *Frictionless contact* (see Table 7).

SUMMARY The proposed structure emerged partially in the data. Concepts *Representing loads at connections* and *Equilibrium* were not identified in any data set. In addition, *Frictionless contact* was not identified in the standard data set, and *Static equivalence* was not identified in the Trad data.

11.3.7 *Summary: Construct*

For establishing construct validity, student interviews were conducted focusing on interpretability of the items. The previously conducted student interviews at pre-instruction level and the expert interviews gave indications for possible pitfalls. Evidence that these indeed affect students' interpretation of the items (and thus construct validity) was found only in some cases, the majority of the observed problems are rooted in lack of understanding of statics instead of mere interpretability issues, which supports claims on construct validity.

The translation analysis revealed several deviations in the translation from the original. Most of the adaptations resulted in more text and complicated phrases. Hence, the German CATS likely tests "reading ability" to a greater extent than the original version, which calls for giving students more time on the test, but they were effectively given less time than in the US administrations. This may explain the slightly weaker performance in comparison to the report by Jorion et al. (2015) on the statistical measures, which are evaluated here by applying the same categorical judgment scheme (see Table 19). The results of this evaluation are described in the following paragraphs and summarized in Table 9.

Table 8: Factor loadings on the factor analysis of the standard data set with eight factors. Loadings of less than 0.30 are not shown. Factors were named according to concepts on the CATS. Concepts marked with an asterisk as well as the concept Equilibrium could not be confirmed. Items marked with an asterisk were excluded from the analysis.

	Drawing forces	Newton's Third Law	Static equiv.	Roller joint	Pin-in-slot joint	Frictionl. contact*	Repres. loads*	Limits on friction
2	0.68							
9	0.47							
15	0.67							
4		0.59						
16		0.56						
24		0.43						
7			0.50					
18			0.36					
27								
1				0.54				
10				0.40				
19				0.58				
3					0.49			
12					0.50			
21					0.36			
8								
17						1.00		
25							0.51	
5								
23							0.57	
14								
6								0.49
13								0.71
22								0.31
11								
20*								
26							0.39	

DIFFICULTY According to the scheme, all item difficulties shall lie between 0.2 and 0.8 for an excellent evaluation of item difficulty statistics. For a good evaluation as obtained by all samples and the reference, up to three items may lie outside of this interval, which is the case here.

DISCRIMINATION As the evaluation of item discrimination is determined only by the worst item in the judgement scheme, the result is only average because of the very poorly performing item 20. Removing this item results in good (samples All, Trad, Tut) to excellent (sample noBlanks) evaluations.

CRONBACH'S α Román et al. (2010b) mention (without reference) that CATS alphas "have fluctuated between 0.70 and 0.90". Similar to Jorion et al. (2015), Steif and Hansen (2006b) report that $\alpha = 0.83$ for the CATS, which is interpreted as a good value for internal consistency. Here, the values for α of the total score are slightly smaller than the reference from Jorion et al. but still in an acceptable (samples All, Trad, Tut) to good range (sample noBlanks) (see Table 6). With only one or two items that decrease the level of internal consistency, the α_{-i} criterium is evaluated as good for all samples which compares to the reference.

IRT In contrast to the reference where the 2PL-model fit best, a 3PL-model was found to better fit the data, i. e. guessing is assumed to be a relevant factor on at least some items. This may be caused by the tighter time constraint, or it may have cultural reasons. The model fit is excellent in case of all samples and the reference.

In the German context investigated here, the test information function peaks strongly towards the higher student abilities. Jorion et al. (2015) report a test information function which has a less pronounced peak and is nearly centered over the average ability level, meaning that in the US context the test measures (1) broader ranges of student abilities similarly well and (2) the average ability students with higher precision than those with very high or low abilities.

STRUCTURAL ANALYSIS The exploratory factor analysis revealed similarities with the hypothesized 9x3 structure but some concepts did not emerge, leading to an average evaluation of the structure. The result improved to a good evaluation when blank responses were ignored. This evaluation compares to the reference for which the handling of blank responses is unclear.

The results by means of the judgement scheme are always at least average, often good or even excellent. Overall, the evaluation is al-

Table 9: Evaluation of analysis according to the categorical judgment scheme in Table 19 proposed by Jorion et al. (2015). E = excellent, G = good, A = average, P = poor, U = unacceptable (did not occur). Numbers in brackets indicate numbers of items which did not comply with the standards. (*corrected error in reference)

	Sample				Reference
Analysis	All	Trad	Tut	noBlanks	Jorion
CTT					
Item stats.					
-Difficulty	G .11 to .71	G .09 to .63	G .12 to .83	G .11 to .71	G* .16 to .78
-Discrimination	A .04 to .45	A .02 to .47	A .02 to .39	A .08 to .69	G* .18 to .65
(-Discr. w/o item 20)	G .15 to .45	G .12 to .47	G .19 to .39	E .27 to .45	-
Total score reliability					
-Cronbach's α total	A .80	A .76	A .79	G .81	G .84
-Cronbach's α_{-i}	G (1)	G (2)	G (1)	G (1)	G (3)
IRT					
Item measures					
(All items fit model)					
-2PL model	-	-	-	-	E (2)
-3PL model	E (1)	E (2)	E (2)	E (1)	-
Structural analysis					
Exploratory FA	A (7)	A (7)	A (7)	G (5)	G (5)

most identical to the reference if blank responses are ignored, and only slightly worse in some aspects for the other samples.

12

SPECIAL INVESTIGATION OF ITEM 20

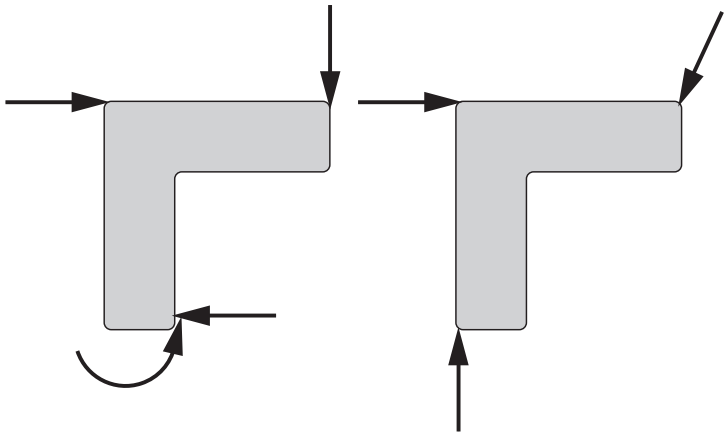
Item 20 (see Figure 27) belongs to the equilibrium concept cluster. Its minimalistic setup successfully sets the focus on the concept. Yet, it was found to be the most problematic item in the quantitative analyses, not only in the German context, but also in the US context (e. g. Jorion et al., 2015; Román et al., 2010b). Nevertheless, evidence has been presented that supports its value in terms of measuring whether students are able to *consistently* invoke *both* force and moment equilibrium. Newcomer and Steif (2008) (and Newcomer and Steif (2007)) analyzed written responses on this item. The question was given to the students twice during a course: once in its original form and once in a slightly modified form by swapping cases (I) and (II) and flipping the L-shaped object and loads horizontally. The given explanations were coded according to whether an equilibrium condition (force or moment) was invoked "(1) never, (2) insufficiently, (3) appropriately, just as needed, or (4) in all cases" (Newcomer and Steif, 2008, p. 488). In addition, the students' awareness or unawareness that forces influence moment equilibrium was monitored. It could be shown that the selected answer on the multiple-choice item tended to match the coding of the explanations, e. g. students who selected (c) tended to produce codes (1) or (2) on force equilibrium, and codes (3) or (4) on moment equilibrium. These results indicate that the item produces valid measurements, despite its high difficulty. Furthermore, US-based results from IRT correspond to the results presented here, "that the item discriminates well, particularly at the high end of the spectrum" (Newcomer and Steif, 2007, p. 3).

The student interviews described in Section 11.3.3 motivated a special investigation of item 20. They revealed that students have no problem interpreting the situation, but it seems as if they just frequently "forget" to check the equilibrium condition for the moments for body (II). This corresponds to the results presented by Newcomer and Steif (2008). If this behavior is due to lack of understanding of equilibrium, the interpretation of the item responses would be valid, unlike if the item is frequently misunderstood.

The interviews gave rise to several hypotheses, why the most attractive distractors C and D are chosen:

1. **Test wiseness**: A point which was also mentioned in the expert interviews is that often one situation is correct and the other one is incorrect. Newcomer and Steif (2008) have acknowledged this notion as "the tacit assumption in typical problems that equilibrium is possible". They furthermore admit that "the style

20. The forces and couple in the two cases shown act at the points indicated. All magnitudes are greater than zero, and the forces and couple act in the directions and senses shown.



Assuming the magnitudes of the forces and couple have the right values, could these bodies be in equilibrium?

(a) I could be in equilibrium; II could be in equilibrium

(b) I could never be in equilibrium; II could never be in equilibrium

(c) I could be in equilibrium; II could never be in equilibrium

(d) I could never be in equilibrium; II could be in equilibrium

(e) Cannot say without more information

Figure 27: CATS item 20

of SCI question studied here - asking whether equilibrium is possible - is certainly different from the typical experience of students."

2. **Distraction**: The students suspected the goal of situation (II) to be testing for understanding that vectors can be decomposed into horizontal and vertical components. The equilibrium condition for moments was "forgotten" over this task.

3. **Check only what failed before**: The equilibrium conditions in two dimensions are habitually checked in a fixed order: $\Sigma F_x = 0$ (horizontal forces), $\Sigma F_y = 0$ (vertical forces), and finally $\Sigma M = 0$ (moments). The equilibrium of body (I) fails already after the second condition. If this condition is fulfilled for the second body, equilibrium is falsely assumed to be possible.

4. **Trigger**: the students often reported after the interviews that the couple in situation (I) triggered them to consider the moment equilibrium condition, while this trigger was missing in situation (II). This speculation has also been articulated by Newcomer and Steif (2007).

It should be noted that none of these hypotheses describe a situation which would suggest that the interpretation of the item may be problematic. If a distractor is chosen for any of these reasons, it must be interpreted as lack of understanding. In addition, testing all of these hypotheses with enough statistical power requires more students than were available for this investigation, therefore, only one hypothesis was tested. Newcomer and Steif (2007) mention that, in their experience, *"students are often content when finding a solution that satisfies one constraint, and they do not exhaustively search for the solution that satisfies all constraints"* . This notion would support the hypothesis "check only what failed before", which was selected to be put to the test.

12.1 METHODS AND IMPLEMENTATION

The hypothesis "check only what failed before" was tested with different variants of item 20 in the last lecture of the 2018/2019 Mechanics I course. The administration was in paper format, and the variants were distributed from multiple evenly mixed stacks of print-outs. As data on the original version is available in abundance from the full test administrations in courses with traditional instruction, the original version was not distributed to make sure that large enough sample sizes on the variants are obtained. The results are therefore interpreted under the assumption that the tested cohort would respond to the original item in a similar distribution as the average traditionally-instructed cohort on the post-test.

To discourage cheating by copying from neighbors, the variants were labelled as eight supposedly different versions, while only three of them were truly different. The given time to complete the task was five minutes, which is more than the average time per CATS item. This decision was made based on the following reasons: Response time data from a more expert-like sample (Direnga et al., 2015b) than the one investigated here suggest that the majority of response times to item 20 lies between 30 seconds and two minutes, and this is after a "warm-up" of seeing 19 similar types of questions. Also, on the entire test, students have the possibility of spending more of their time on items which are difficult for them and to save time on easy items. This possibility is not given on a one-item test, which is why allowing more than the average time seems appropriate.

Two of the three different variants were designed to check the above-mentioned hypothesis. While the question format is still closed-ended, both variants 1 and 2 offer room for notes. Additionally, one variant was suggested by the instructor with the aim of adapting the item more to the style of questions the students are used to (see Figure 49).

- Variant 1 (versions labelled 1/2/3): The sequence of cases (I) and (II) is swapped. If the original case (II) is now investigated first as case (I), "check only what failed before" now applies to the sum of moments which should be checked in both cases more frequently, while the force equilibrium conditions are expected to be less often or - because of the theorized fixed sequence of checking the conditions - equally often applied in which is now case (II). A shift from distractor (d) to distractor (c) (ignoring the swap) or a higher frequency of the correct response (b) is thus considered as an indicator for accepting the hypothesis.

- Variant 2 (versions labelled i)/ii)/iii)): The vertical force on body (I) was removed and the senses of both horizontal forces were reversed, resulting in a possible equilibrium if considering only forces, but not when moments are considered. If the hypothesis is true, the sum of moments should be checked more frequently for body (II) as well, and the frequency of correct responses should increase compared to the original.

- Variant 3 (versions labelled A/B): This instructor-designed variant was not designed to test the hypothesis. While no changes were made to the loads acting on the body, this variant provides slightly more guidance: The arrows are labelled, students are encouraged to introduce own dimensions if needed, and the final judgement prompt hints again at the conditions that only positive values for the forces and moment are allowed. The question is semi-open-ended in that it requests the students to show their

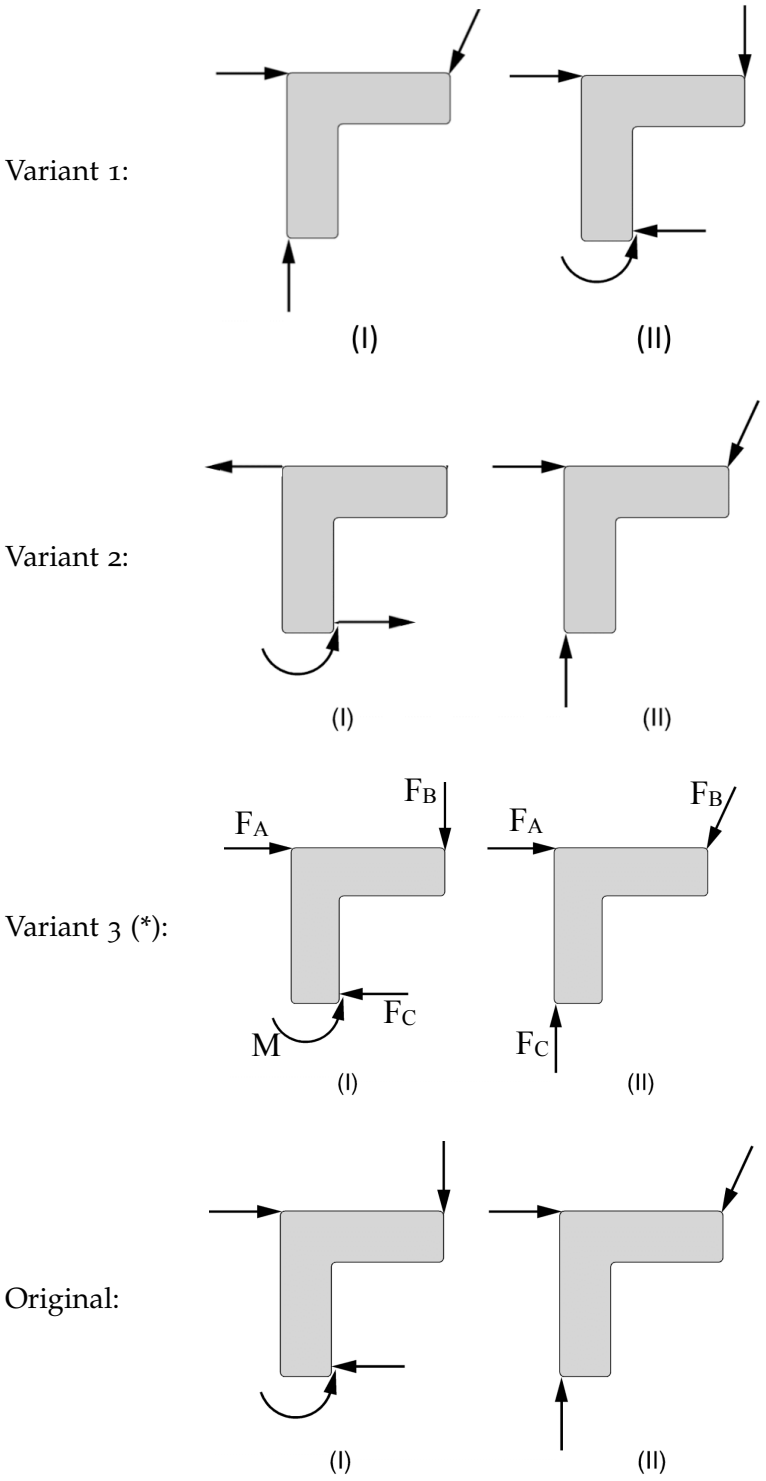


Figure 28: Different versions of item 20. (*) Variant 3 differs from the original in more aspects than shown here (see Figure 49 for the complete task).

work, but asks for a final judgement. The response for case (I) is clearly separated from the one for case (II), avoiding complicated combinations of judgements as in the original response options.

The response distributions on the variants are compared to the one on the original item 20 from the traditional instruction sample. If the response distributions differ strongly, the aspect which was addressed in the variant probably has an effect.

12.2 QUANTITATIVE RESULTS

The sample sizes were 118 for variant 1, also 118 for variant 2, and 49 for variant 3. The response frequencies on the variants that test the hypothesis are shown in Figure 29, rounded to the nearest 5 %. (As variant 3 was not created to test the hypothesis, the results are reported in the appendix, see Figure 48.) First of all, it is worth noting that there *are* differences among the variants and between the variants and the original. It seems that even subtle changes like swapping the sequence of the inspected cases are not only superficial. As described above, Newcomer and Steif (2008) used the item with superficial changes, assuming that the item remains comparable to the original. The results presented here suggest that their assumption may be incorrect and therefore their conclusions considering the change in the students' conceptions over the course may be invalid.

For variant 1, the frequency of correct responses is comparable to the original, but the distribution of false responses differs. There seems to be a "shift" from distractor (a) to distractor (d). Thus, more students recognize the non-zero resultant vertical force than on the original question, but the equilibrium condition for moments was not considered more often in the critical case. Thus, based on the data from variant 1, the hypothesis is likely to be rejected.

Variant 2 shows a substantially higher frequency in the correct response. Also, the frequencies of distractors (a) and (c) are strongly reduced in comparison to the original, indicating that, overall, students recognize more easily that equilibrium is impossible in the modified case (I). Yet, a remarkable number of students still claim that the body in case (II) can be in equilibrium. Thus, based on the data from variant 2, the hypothesis is likely to be rejected.

The data collected from both variants 1 and 2 of item 20 contradict the hypothesis that the equilibrium condition of moments is more often applied to case (II), if the equilibrium in case (I) was found to fail due to the sum of moments.

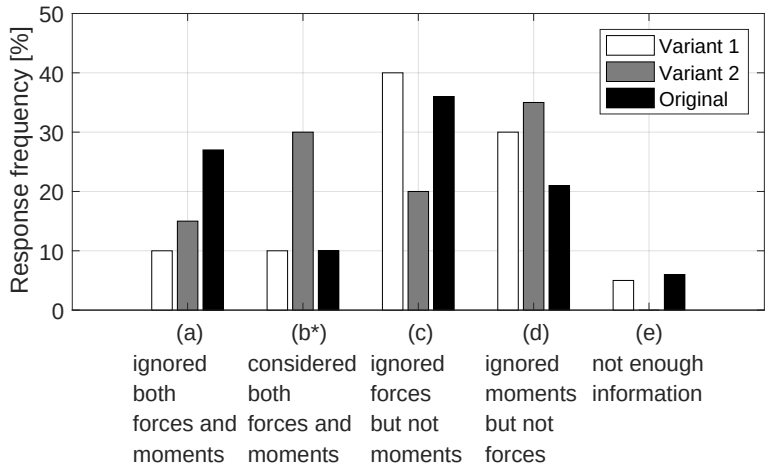


Figure 29: Answer distributions differ on variants of item 20. For variant 1, the bars representing response options (c) and (d) have been swapped here in accordance with the cases to facilitate comparability. The correct response is (b) as marked by the asterisk.

12.3 QUALITATIVE RESULTS

Inspection of the available notes and the open-ended variant 3 allowed for a qualitative view on the data and revealed a new hypothesis, what most students struggle with.

HYPOTHESIS: It is often not recognized that a *non-positive* result for a force or a couple must lead to the conclusion that equilibrium cannot be obtained.

The condition that the drawn forces and couples must not be zero is more often noticed and understood than the condition that negative values are not allowed. The passage in the item stem "ungleich Null" ("unequal to zero") is more often marked by the students than the passage "Drehsinn und Richtungssinn [...] zutreffend wiedergegeben" ("the forces and couple act in the directions and senses shown"). Out of 24 students who marked any text in the item stem, 17 students marked "ungleich Null", 1 student marked the passage on the shown directions and senses only, and 3 students marked both. The English version speaks of "All magnitudes are greater than zero". In the modified variant 2, this subtle difference introduced by the German translation becomes important, because negative values can lead to equilibrium of the modified case (I). As this is not the case for the original variant, the CATS results are unaffected by a hypothetical misunderstanding introduced by this translation.

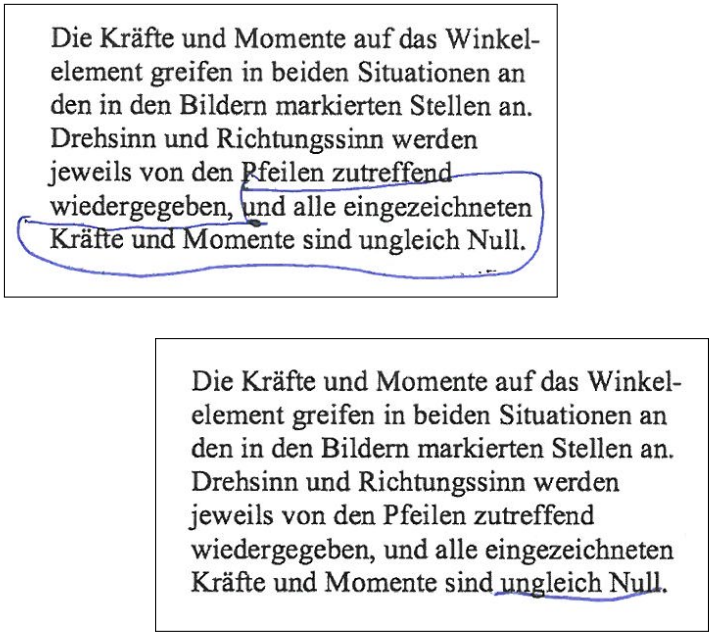


Figure 30: Examples of the markups made by students in the stem of item 20. The passage "alle eingezeichneten Kräfte und Momente sind ungleich Null" ("all shown forces and moments are unequal to zero") is often marked by the students but not the passage "Drehsinn und Richtungssinn werden [...] zutreffend wiedergegeben" ("the forces and couple act in the directions and senses shown").

12.4 CONCLUSION

The hypothesis that the equilibrium condition of moments is more often applied to body (II), if the equilibrium of body (I) was found to fail due to the sum of moments, *cannot* be confirmed by the data. Therefore, no additional insight on why this item is so overly difficult was gained. However, it was found that students often do not realize that a *non-positive* result for a force or a couple obtained by solving the equilibrium equations must lead to the conclusion that equilibrium cannot be obtained in the given situation. This issue, however, does not affect the CATS results or its validity.

13

DISCUSSION AND CONCLUSION OF THE REVALIDATION STUDY

Validation is the process of collecting and interpreting evidence that the interpretation of test results are valid in the given context and for the defined purpose. Evidence was presented that addressed different aspects of validity: content and face, criterion, and construct.

CONTENT AND FACE The analysis of textbooks and course descriptions shows that the content of the CATS is covered by instruction. The concepts selected by the test developers were largely confirmed by the interviewed experts as the most central ones to statics instruction. Criticism involved the strong focus on specific connections such as rollers and pin-in-slot joints. Instead, the results from the expert interviews and textbook analysis indicate that introducing items on the concept of static determinacy could further enhance content validity in the German context.

Other concepts perceived by the experts to be central and missing are *center of gravity and centroid*, *internal forces and moments*, *structural analysis*, *virtual work*, and *determining reaction forces*. As the CATS has a limited scope and the included concepts were likewise assessed to be central, adding these concepts to the CATS would result in other cutbacks, such as less precise measurements or more time on the test. Such a modification would hence not necessarily lead to a higher quality instrument.

CRITERION Correlations with instruments measuring similar constructs (midterm and final exams) were found to mirror results from the US context. Likewise, the dominant distractors are largely the same, indicating that the basis for the instrument development (typical student errors) is valid for both national contexts.

CONSTRUCT The interviews revealed that most (but not all) items are correctly interpreted by students and experts. Many elements on the items that pre-instruction students struggle to interpret are clear to the target population of post-instruction students such as the terms "moment" or "free-body diagram", but some issues remain even after instruction, such as the misinterpretation of arrows as indicating movement instead of forces or the interpretation of "arbitrary forces". Additional interpretation issues were fostered by instruction such as not interpreting single arrows as complete two-dimensional force vectors because students are trained to think in force *components*.

The student interviews were mainly conducted to probe for interpretability of the items and not for correct student reasoning. One issue in terms of student reasoning still emerged from the interviews that affects the validity of the Newton's Third Law concept: It is unclear whether students selecting the correct response do so because they correctly applied Newton's Third Law or because they incorrectly reason that the equilibrium condition of forces ($\sum \vec{F} = 0$) applies to the two forces shown in the correct response option. A two-tier item design asking the students to select the appropriate reasoning could improve the diagnostic quality of this concept category.

Applying the quantitative framework by Jorion et al. (2015) reveals that the most problematic item is the same in the German and the US context, and that the German CATS data shows a slightly weaker but still reasonable performance on structure, reliability, and discrimination than the reference. The range of the most precise measurement is rather ill-matched to the student population. Possible reasons might be that the German students are less familiar with the representation, that the translated text is more complicated than the original, or that less time was given.

IS THE TIME LIMIT REALLY AN ISSUE? Anderson et al. (2009) administered a computer-based CATS in the US without time limit which the majority of the students completed within 30 minutes. While this is also true for the data discussed in this dissertation - the majority of students (75 %) finish in time - the CATS is not supposed to be a power test, meaning that the time required to finish the test should not play a role in the interpretation of the test result. The increasing number of blank responses towards the end of the test may result from lack of time but also from lack of motivation or engagement. The results from the textbook and translation analyses speak for lack of time as they suggest that German students may require more time than US students because of the unfamiliarity with the graphical representations and the demand for better reading comprehension skills. In addition, the time limit was frequently criticized as too short by the interviewed experts, but literature suggests that they may overestimate the time required by the students: "Because experts attempt to understand problems rather than to jump immediately to solution strategies, they sometimes take more time than novices (e.g., Getzels and Csikszentmihalyi, 1976)" (Bransford et al., 1999, p. 44). Impressions from the student interviews suggest that lack of time is indeed an issue for some students, but an interview situation cannot be compared to a test administration, so the main reason for the pattern of blank responses cannot be resolved here. In any case, the blank responses on the late items are problematic for the interpretation of the data, and filtering the data so that no assumptions need to be made on the meaning of a blank response introduces another sampling bias.

As shown above, the analysis based on such a subsample leads to already slightly better results in the evaluation scheme of Jorion et al. (2015), but allowing more time could possibly further improve the quality of the measurement for the following reason: Assuming that the allowed time on the test drives the test-taking behavior, the abilities of students who adapt their speed according to the limit are probably underestimated because the students make mistakes that they would not make if given enough time. Similarly, slow-working individuals might have the proficiency to correctly respond to late items, but they are not given the chance, resulting again in underestimation. Consequently, late items with a large frequency of blank responses are overestimated in difficulty, independent of whether student data including blank responses are considered or omitted. With more time available, the test information function is therefore expected to better match the ability levels of the population, provided that students remain motivated and engaged for the duration of the test. The precision of the test could thus be improved by allowing more time. The question is how much more. Since nearly all students (independent of whether they come from Trad or Tut cohorts) respond to 15 out of 27 items in 30 minutes, one may linearly extrapolate the required time for 27 out of 27 items to 54 minutes. This time can be taken as a conservative estimate, as Steif and Hansen (2007) found "no noticeable pattern in the variation of mean scores for the remainder of examinees with time above 25 minutes". Even though these results stem from the US context and with unlimited time allowed, they indicate that if there is an effect of giving substantially less time on the test compared to the suggested hour it may be less serious than expected. Increasing the allowed time to 40 or 45 minutes could thus already increase the quality of the measurements significantly, while keeping the use of class time low.

CONCLUSION Overall, the presented evidence allows to conclude that the CATS total score can be interpreted as a measure for conceptual understanding of statics for the intended purpose of assessing the effectiveness of learning materials in the German higher education context. This is especially true for interpretation of aggregated and averaged data instead of individual scores as the influence of measurement errors is reduced by two processes: aggregating responses to a total score and averaging student total scores. Interpreting student understanding of the concept sub-scales, even at the cohort level, is only partially valid and should be done selectively. A collection of suggestions for improvement of the German CATS is given in Appendix G.

Part II

METHODS TO EVALUATE AND COMPARE RESULTS FROM PRE- AND POSTTESTS

"[F]or many subjects in engineering, while there are certainly concepts in previous courses that are relevant, a test that measures conceptual development adequately by the end of the course may not be a valuable measure at the beginning of the course."

(Steif and Hansen, 2007, p. 209)

14

INTRODUCTION TO THE ANALYSIS METHODS STUDY

The previous part was dedicated to the validation of score interpretation from an existing concept inventory for use in the German higher education statics context. This part discusses statistical methods to evaluate concept inventory scores when used in a pre- and post-test setting. Specifically, the focus lies on the evaluation of scores from non-identical pre- and post-tests (NIPPs), i. e. when the instrument used as post-test is different from the one used as pre-test. Ideally, the pre-instruction understanding is measured with the same test instrument as the post-instruction understanding. This research design has many advantages, such as allowing for direct comparison of total scores as well as examination of shifts in responses to individual items. An essential prerequisite is the existence of a test instrument which allows for valid score interpretations of both measurements, pre and post. Such an instrument must fulfill three criteria to produce valid results:

1. It must be interpretable for students at pre-instruction level.
2. It must test understanding of the concepts subject to the course.
3. It must be scaled such that neither measurement, pre or post, produces significant floor or ceiling effects[1].

Development of such an instrument is challenging, and even though the number of available standardized test instruments has increased strongly during the past two decades, these criteria are often not met for the following reasons.

1. Courses generally introduce new concepts and vocabulary. In order to test student understanding of the concepts taught, it is often necessary to use this new vocabulary, which leads to problems with pre-instruction interpretability. For instance, pre-instruction mechanics students have some conception of "friction" or "force" (whether correct or incorrect), but they usually have no idea what is meant by "free-body diagram", "virtual work" or even "moment" or "torque" (Geier, 2016). As long as the test allows for a response that accurately reflects this non-existent conception, (e. g. a "no idea" response option), the test might yield valid results. Otherwise, students are forced to

1 High frequency of extreme scores on the lower (floor) or upper (ceiling) end of the scale. Indicates that the scale is inadequate for the population.

guess which dilutes the validity of other responses made with high confidence. Even worse, in case unfamiliar terms or symbols are used in items that aim to probe the understanding of a possibly familiar concept, students may not be able to respond according to their actual understanding.

2. The score range of any test instrument is finite. Strong occurrences of floor or ceiling effects have a negative impact on the validity of test score interpretations. Score distributions that are heavy on either one of the range extremes should therefore be avoided by choosing appropriate instruments. Due to instruction between the tests, an improvement in understanding is expected. On identical pre- and post-tests (IPPs), this would be represented by a shift in score distribution towards the higher end. For large effects of instruction or if the pre-test scores are already high, this can lead to the problem that the real shift in understanding cannot be displayed by the limited score range.

Despite these difficulties, there are examples of successful IPPs study design. The FCI is one of the few instruments that has been shown to work well as IPPs in introductory physics courses. Its core concept (force) is present in everyday life which fosters the construction and reinforcement of an unscientific mental model based on everyday observations (e. g. "Moving objects require a driving force, otherwise they stop, for example a rolling ball.", "In an accelerating car, the seat exerts a force on the passenger, who is relaxed and therefore does not exert any force on the seat.", "Heavy objects fall faster than light ones, for example a rock vs. a feather."). The item design exploits such everyday examples. Hake (1998) described the FCI as "understandable to the novice who has never taken a physics course, while at the same time rigorous enough for the initiate".

In case a single instrument does not meet all the criteria, using two different but related instruments is an alternative approach. One suitable test pair for introductory engineering mechanics is the FCI and the CATS. These instruments are related since a thorough understanding of forces is required to do well on either test. Despite being non-identical scales, a good performance on the first can be seen as indicative of a good basis for acquiring the skills for the second. The FCI hence fulfills the purpose of a baseline measurement for the CATS. Part III of this dissertation discusses results from a series of pre-/post-testing in a mechanics course with the FCI as pre-test and the CATS as post-test. In this part, methods to analyze such data are investigated as the use and evaluation of NIPPs has not yet become a standard in the STEM Education Research communities. While the established evaluation methods for IPPs have been thoroughly discussed (e. g. Cronbach and Furby, 1970; Hake, 2002; Bond, 2005; Marx and Cummings, 2007; Hake, 2010), evaluation methods for NIPPs have not

been addressed to the same extent. Analysis of covariance (e. g. Engqvist, 2005) and multiple linear regression (e. g. Theobald and Freeman, 2014) have been proposed, but are often not suitable as will be explained below.

Parts of this investigation have been published in Direnga et al. (2014, 2017, 2018). First, shortcomings of IPPs are discussed and arguments why using NIPPs is often a better option are brought forward. An overview of established methods for evaluating pre- and post-test data is presented. As a solution for analyzing NIPPs, the Discriminative Learning Gain (DLG) is then introduced. Based on linear regression with confidence bounds, it serves to quantify as well as easily visualize the difference in courses with respect to their performance on NIPPs. While the raw test scores contain information at the individual student level, applying a regression reduces the complexity of the data to two parameters which describe course performance. This reduction of complexity is necessary to effectively quantify differences in effectiveness between courses. The interpretation of the resulting parameters will be discussed in detail below. Additionally, the application of an effect size measure to the regression model is demonstrated. A comparison of the methods is made.

15 ESTABLISHED METHODS

In this chapter, an overview of some established methods for evaluating learning success will be presented: the average normalized gain (g), normalized change (c), and analysis of covariance (ANCOVA). An overview of similarities and differences among these methods and the Discriminative Learning Gain (DLG), which will be presented in Chapter 16, can be found in Table 11 in Chapter 17.

15.1 AVERAGE NORMALIZED GAIN

Using some sort of gain or change measures is quite common in PER and EER. The difference between the pre- and post-test scores gives insight on what students have learned in absolute terms. For direct comparisons of such absolute gains from courses with very different average pre-test scores, a normalization is required to consider possible ceiling effects. Otherwise, courses with low pre-test scores have a greater chance of achieving a large absolute gain than those with high pre-test scores, simply because "there is still more to gain".

Hake (1998) conducted a historically important study (Beichner, 2009) that uses such a normalized learning gain to compare interactive-engagement and traditional teaching in various physics classes. The average normalized gain, denoted by g, is a well-established measure for assessing data on teaching effectiveness (e. g. Girwidz et al., 2003; Finkelstein and Pollock, 2005; Andrews et al., 2011). It is defined as follows (Hake, 1998):

$$g = \frac{\text{absolute gain}}{\text{maximum possible gain}} = \frac{(\text{post})_{\text{ave}} - (\text{pre})_{\text{ave}}}{100\% - (\text{pre})_{\text{ave}}}. \tag{15}$$

The absolute gain is the gain of class average pre- and post-test scores, and $(\text{pre})_{\text{ave}}$ and $(\text{post})_{\text{ave}}$ are the class average scores of the pre- and post-tests, respectively, given as a percentage of the maximum possible score.

Lines of constant g can be visualized by plotting the absolute gain vs. the average pre-test score, as shown in Figure 31. The range of g starts at the horizontal ($g = 0$) and reaches its maximum of $g = 1$ at a slope of -1. It should be noted that the special case of $\text{pre} = 100\%$ is not mathematically defined by Equation (15), but since this case is highly unlikely due to several factors (operation on averages, good test construction) and easy to interpret (no more gain possible), there is no need for further elaboration.

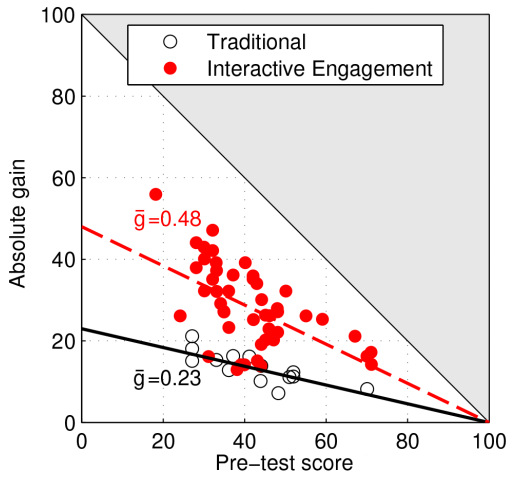


Figure 31: Illustration of the average normalized gain based on data from Hake (1998).

The data displayed in Figure 31 was taken from Hake (1998) for illustration purposes. Each datapoint displayed corresponds to one course. If linearity of the measure is assumed, it can be seen that the average g of the interactive-engagement courses is slightly more than twice as high compared to the average g of the traditional courses. Hake's study and the quantification of learning through g give instructors the opportunity to compare their achieved gain to these values, independent of the average class pre-test scores and the administered test instrument.

Still, there have been critical voices by Marx and Cummings (2007) concerning the definition of g. Based on the reasonable assumption that the performance on an identical test will improve after instruction[1], the definition stated above was designed for positive gains, only. Although it would not be mathematically incorrect to apply Equation (15) for post $<$ pre, g-values in the range $[-\infty, 0)$ can be attained, which makes averaging more problematic. Also, the interpretation is questionable when relating a loss to the maximum possible gain. A loss at an already low initial score would be considered less negative than the same absolute loss at a higher initial score, which is contrary to the intent of the g. Marx and Cummings (2007) proposed a different definition for negative values which is presented in the following section.

15.2 NORMALIZED CHANGE

To make g applicable to negative gains in an easy-to-interpret manner, the normalized change described by Marx and Cummings (2007) can

1 While this is especially true for class averages, the chances of obtaining negative gains are higher if the measure is applied to individual scores.

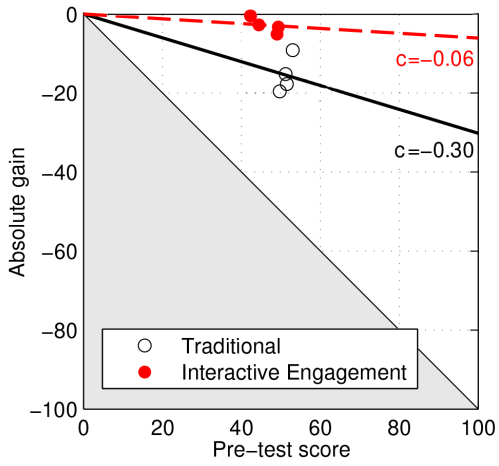


Figure 32: Illustration of the normalized change, showing data from non-identical pre- and post-tests.

be applied. While a positive absolute gain is related to the maximum possible gain as in Equation (15), a negative absolute gain is related to the maximum possible loss.

$$c = \begin{cases} \frac{\text{post}-\text{pre}}{100\%-\text{pre}} & \text{if post} > \text{pre} \\ \\ \frac{\text{post}-\text{pre}}{\text{pre}} & \text{if post} < \text{pre} \end{cases} \tag{16}$$

Note that Equation (16) represents the normalized change of *average* scores, and thus differs from the definition given by Marx and Cummings (2007) who advise to first calculate the normalized change of individual students and then calculate the average. This procedure requires the data from pre- and post-tests to be linked. Marx and Cummings (2007) also state that "for large numbers of students the numerical difference is small" between averaging the scores and averaging the gains. Under these circumstances, c could still be applied to averaged unlinked data without having to expect larger errors. For the comparative purposes in this part of the dissertation, the averaging procedure used for g will be pursued for c as well.

The diagram shown in Figure 32 can be seen as an extension of Figure 31. The data[2] was collected using the FCI and CATS as NIPPs. All c are negative. In case of IPPs this would be interpreted as "unlearning" or systematic creation of misconceptions. In case of NIPPs, this could also indicate that the post-test was more difficult than the pre-test. It can be seen that the normalized change in the interactive-engagement courses, $c_{IE} = -0.06$, is greater than in the traditional courses with $c_T = -0.30$. This corresponds to the results found by Hake (1998).

Even though the definition of c allows a reasonable interpretation of negative gains and comparisons between gains, these are only

2 For details, see Direnga et al. (2014).

valid when comparing gains from the same pre-/post-test combination. This is due to the fact that a gain of zero cannot be interpreted as "no learning", as the pre- and post-test had a different level of difficulty. Therefore, the data shown in Figure 31 is explicitly not displayed here together with the NIPPs data as this would suggest that a comparison between these datasets was valid. However, it is feasible to compare the *differences* in average gains Δc for any two courses, i. e. for interactive engagement and traditional courses:

$$\Delta c_{IE-T} = c_{IE} - c_T. \tag{17}$$

If Equation (17) is applied to the data displayed in Figure 31 (Hake's data from IPPs) and Figure 32 (the data from FCI/CATS NIPPs), respectively, the results are similar.

$$\text{Figure 31:} \quad \Delta c_{IE-T} = 0.48 - 0.23 = 0.25 \tag{18}$$

$$\text{Figure 32:} \quad \Delta c_{IE-T} = -0.06 - (-0.30) = 0.24 \tag{19}$$

Assuming that instruction using interactive-engagement does result in a greater learning effect and that the employed NIPPs combination does measure this effect, this similarity supports the hypothesis that the normalized change is a valid statistical method for assessing NIPPs data.

General criticism of measuring score change has been expressed by Cronbach and Furby (1970). They argue that, although gain or change measures are widely used, they are not appropriate or even necessary in most cases. Instead, they recommend regression-based analysis methods such as ANCOVA to be used. Their article "How We Should Measure 'Change' - Or Should We?" was commented by Bond (2005), Hake (2010), and many others. In Hake's opinion, this "pre/-post paranoia" is one of the reasons that results from education research often do not have an impact on reforms.

Another disadvantage shared by g and c is a possible bias in favor of high pre-test populations. Suggesting the existence of such a bias, Coletta and Phillips (2005) found a linear relationship between gain and pre-scores when using individual student scores and gains. In contrast, Von Korff et al. (2016) found no correlations of such kind when operating on class averages. They conclude that gain is a "powerful tool" for the purpose of comparing classes this way. Nissen et al. (2018) criticize g for the assumption that it is more difficult to gain scores for students with higher pre-test scores in comparison to those with lower scores, and thereby not considering that knowledge or understanding is gained more easily with more prior knowledge (Bransford et al., 1999). They suggest to use for example Cohen's d as a measure of effect size instead.

15.3 ANALYSIS OF COVARIANCE

In non-randomized studies, differences in treatment groups with respect to concomitant variables (e. g. pre-test scores) can obscure the visibility of the treatment effect (e. g. instruction) on the outcome variable (e. g. post-test scores) under investigation. Analysis of covariance (ANCOVA) has proven to be a powerful and widely used tool to account for such variables.

The model for a single-factor ANCOVA with fixed treatment effect (i. e. the effect of instruction is assumed to be independent of the pre-test score) can be written as (Neter et al., 1985, pp. 848-849)

$$Y_{iq} = \bar{Y}_{..} + \tau_q + \gamma \left(X_{iq} - \bar{X}_{..} \right) + \varepsilon_{iq}, \tag{20}$$

where

- X_{iq} and Y_{iq} are the observed pre- and post-test scores of the i-th individual on the q-th type of instruction,
- $\bar{X}_{..}$ and $\bar{Y}_{..}$ are the mean pre- and post-test scores over all instruction groups,
- τ_q are fixed instruction effects with the convention that $\sum \tau_q = 0$,
- γ is a regression coefficient denoting the effect of X on Y (i. e. the slope), and
- ε_{iq} are assumed to be independent and normally distributed errors.

Considering that the expectation value of the random error is zero ($E(\varepsilon_{iq}) = 0$) and the one of the mean post-tests is some constant ($E(\bar{Y}_{..}) = \mu$), regression lines for each instruction type are expressed as

$$\mu_{iq} = \mu + \tau_q + \gamma \left(X_{iq} - \bar{X}_{..} \right). \tag{21}$$

It is evident that γ, the regression coefficient representing the slope, is assumed to be independent of the instruction type q. This assumption is not always adequate, and a failure to recognize when it fails can lead to a misinterpretation of the results (Engqvist, 2005). This is because the significance tests performed by ANCOVA tools test for difference in intercepts. Assuming equal slopes, the result of this test can be translated to any other value of the concomitant variable. If this assumption is inadequate because a model allowing for different slopes with converging or diverging regression lines (which may even cross within the range of the pre-test score scale) yields a better representation of the data, the result of the significance tests may only be interpreted for the intercept $X_{iq} = \bar{X}_{..}$. A significant difference

in the intercepts then does *not* allow to conclude that the regression lines differ significantly for *all* values of the concomitant variable. In this case, literature suggests that *"[w]hen the treatment regression lines interact with the concomitant variable in the form of nonparallel slopes, covariance analysis is not appropriate. Instead, separate treatment regression lines should be estimated and then compared" (Neter et al., 1985, p. 851).*

16

DISCRIMINATIVE LEARNING GAIN (DLG)

Following this overview of some established methods, the DLG is proposed as a tool to compare the effectiveness of teaching based on pre- and post-test score pairs. When ANCOVA fails due to non-parallel slopes, Neter et al. (1985) suggest that "separate treatment regression lines should be estimated and then compared". The DLG is based on a regression line with confidence bounds. Here, the pre-test scores X_i are the independent variable and the post-test scores Y_i are the dependent variable. As the regression lines are estimated separately, the index q indicating the different instruction groups will be neglected for better readability. An example of the raw data, namely the distribution of pre- and post-test scores for one group, is shown in Figure 33. Even though the variation is rather large in the individual data, the mean post-test scores $\bar{Y}_j$ at each possible pre-test score level X_j are quite well described by a linear model (see Figure 34). As mentioned above, only inferences about an entire course are to be drawn instead of about individuals. Therefore, the linear model is appropriate even for large variation in individual data.

16.1 REGRESSION LINE

The linear model is given by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{22}$$

The unknown line parameters of the model, β_0 and β_1, are estimated by an ordinary least squares approach by minimizing the following term with respect to the variables β_0 and β_1:

$$Q = \sum_{i=1}^{N} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2. \tag{23}$$

The result are the point estimators b_0 and b_1

$$b_1 = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}, \tag{24a}$$

$$b_0 = \bar{Y} - b_1 \bar{X}, \tag{24b}$$

where $\bar{X}$ and $\bar{Y}$ are the overall average pre- and post-test scores. So far, the model was discrete with the indices i and j denoting the individual observations of discrete test scores and the distinct pre-test score levels, respectively. In the following equation, which makes use

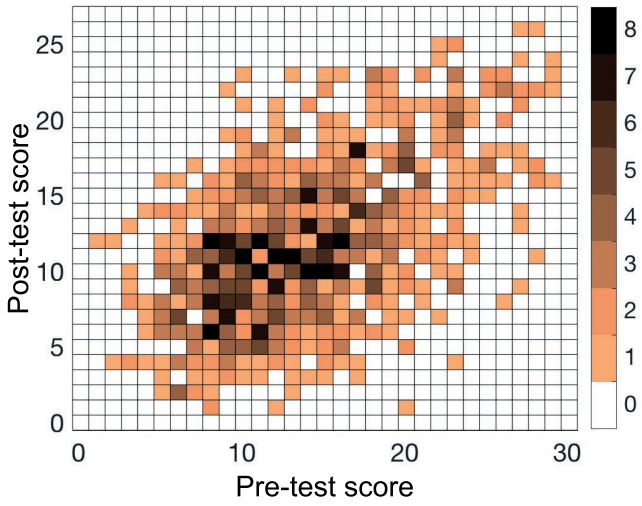


Figure 33: Distribution of individual score pairs on non-identical pre- and post-tests, N = 828.

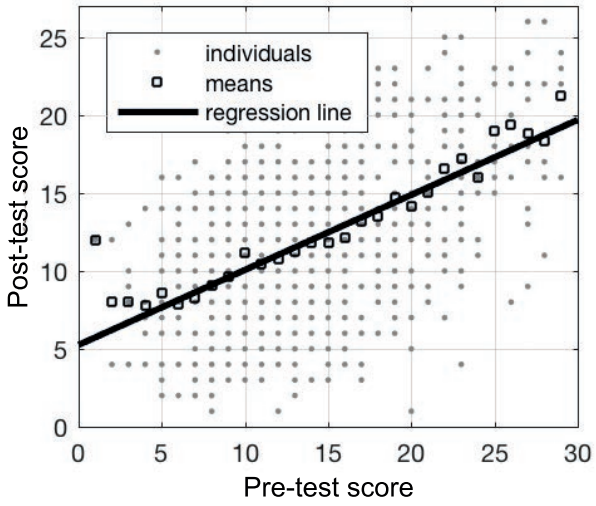


Figure 34: Mean post-test scores for each pre-test score and resulting regression line. (Multiple occurrences of identical score pairs are not represented here, see Figure 33 instead.)

of the point estimators, the index h denotes that the variables X and Y are now assumed to be continuous on the pre- and post-test scales, respectively:

$$\hat{Y}_h = b_0 + b_1 X_h. \tag{25}$$

The line is best interpreted as the expected post-test score for "model students" with the respective pre-test scores, given that linearity is assumed and the best fit to the data is applied. The DLG provides two valuable pieces of information:

1. The general post-instruction level is represented by $\hat{Y}_{50\%}$, the estimated post-test score at a pre-test score $X_{50\%}$ of 50 %, which is equivalent to the average value of the line.
2. The discriminative effect of the instruction for students with different pre-test score levels is represented by the slope b_1. It shows how subgroups of students with different pre-instruction levels responded to the instruction.

Because of the definition of the general post-instruction level as the estimated value at $X_{50\%}$, one may prefer to shift the model such that the general post-instruction level is the intercept. Note that the illustrations in this chapter are all based on the unshifted model, but for sake of completeness, the shifted model description, analogous to Equations (22) - (25), is given as:

$$Y_i = \beta_0 + \beta_1 (X_i - X_{50\%}) + \varepsilon_i, \tag{26}$$

$$Q = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 (X_i - X_{50\%}))^2 \tag{27}$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}, \tag{28a}$$

$$b_0 = \bar{Y} - b_1 (\bar{X} - X_{50\%}), \tag{28b}$$

$$\hat{Y}_h = b_0 + b_1 (X_h - X_{50\%}). \tag{29}$$

A high general post-instruction level as well as a non-negative slope are of course favorable. A large positive slope would show that the stronger model student benefits a lot more from instruction than the weaker model student, while a negative slope would indicate that

either the tests might not be valid in a way that they do not measure the intended criteria, or the instruction created confusion which rather affected students with good conceptual understanding at the beginning of instruction. A slope close to zero represents an equalizing effect of instruction. Whether the ideal course should have a strong positive slope (i. e. students with a significantly higher pre-test score should also have a significantly higher post-test score compared to low scoring students) or not (i. e. students should reach more or less the same level after instruction independent of their pre-test scores) is rather a matter of the personal view on education.

What the DLG generally *cannot* do (just like the established methods) is predict individual students' post-test scores from their pre-test scores because of the often large variations in individual scores (see Figure 33). Furthermore, the resulting DLG line parameters depend strongly on the particular pre- and post-tests used. To illustrate this fact, suppose a group was randomly divided into two subgroups. They take the same post-test but different pre-tests. Because the two subgroups were formed randomly, they have the same ability such that one subgroup scores high on the easier pre-test while the other subgroup scores low on the more difficult one. For the same reason, the score distribution on the post-test should *not* differ between the two subgroups if the instrument is the same. This can be illustrated as a horizontal shift of the data in Figure 33 and consequently in Figure 34. The general post-instruction level of the subgroup taking the easier pre-test will therefore be lower in comparison to the other subgroup, which would be an incorrect conclusion.

NIPPs generally do not allow to identify the pre- and post-test score pairs that are the equivalent to "no learning", at least not without a prior calibration study on the specific test pair[1]. Therefore, a single regression line does not inform about the effectiveness of the instruction. Instead, it must be compared to other data sets using the same test pair. When doing this, statistical uncertainties need to be considered. One way to take these into account is to determine the confidence bands around the regression lines, as shown in the next section.

16.2 CONFIDENCE BOUNDS

When dealing with sampled data, the resulting regression parameters, both slope b_1 and intercept b_0, can only be estimated with a certain confidence $1 - \alpha$. By using a joint estimation technique, one

1 Such a calibration would be conducted by giving both tests at pre-instruction level to the same student sample (see Thornton et al. (2009) for the test pair FMCE/FCI), however, if NIPPs are the method of choice, it is likely that the results on the post-test instrument at pre-instruction level are not meaningful. Otherwise it could have been used as IPPs.

can exploit the fact that β_1 and β_0 are not independent. The inequality

$$\frac{N(b_0-\beta_0)^2+2\left(\sum X_i\right)(b_1-\beta_1)+\left(\sum X_i^2\right)(b_1-\beta_1)^2}{2\,MSE}\ldots$$
$$\ldots\leqslant F(1-\alpha;2,N-2) \quad (30)$$

describes an elliptical $1-\alpha$ confidence region in the β_0,β_1-space (Neter et al., 1985, pp. 147-148). Here, MSE is the mean squared error

$$MSE=\frac{\sum\left(Y_i-\hat{Y}_i\right)^2}{N-2}, \quad (31)$$

Furthermore,

- N is the sample size,
- α is the confidence coefficient, and
- $F(1-\alpha;2,N-2)$ denotes the inverse of the F-distribution with 2 degrees of freedom in the numerator and $N-2$ degrees of freedom in the denominator for the percentile $1-\alpha$.

All β_0,β_1-pairs contained in this region determine the set of possible regression lines. From this, a hyperbolically shaped confidence band around the estimated regression line delimited by the upper and lower confidence bounds can be derived:

$$c_{upper,lower|h}=\hat{Y}_h\pm W\cdot\left|s\left(\hat{Y}_h\right)\right|. \quad (32)$$

$\hat{Y}_h$ is defined by Equation (25), W is a constant given by

$$W^2=2F(1-\alpha;2,N-2), \quad (33)$$

and the standard deviation of the estimated points is

$$s\left(\hat{Y}_h\right)=\sqrt{MSE\cdot\left[\frac{1}{N}+\frac{\left(X_h-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}\right]} \quad (34)$$

(Neter et al., 1985, pp. 154 and 74). This results in the following expression for the confidence bounds:

$$c_{upper,lower|h}=\hat{Y}_h\ldots$$
$$\ldots\pm\sqrt{2F(1-\alpha;2,N-2)\cdot\frac{\sum\left(Y_i-\hat{Y}_i\right)^2}{N-2}\cdot\left[\frac{1}{N}+\frac{\left(X_h-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}\right]} \quad (35)$$

Note that the values of the variance for any h also depends on data from other pre-test scores.

The regression line parameters b_0 and b_1 given by Equation (24) can also be obtained by performing a *weighted* least squares approach on the means (see e. g. Neter et al., 1985, p. 167). In this case, the weights are given by the frequency of each pre-test score X_j. Direnga et al. (2014) and Direnga et al. (2017) have focused on this equivalent approach referred to as weighted linear regression (WLR). The confidence bounds, on the other hand, must be calculated based on the individual data. By considering only the weighted means, one would lose valuable information about the confidence because the number of points that the confidence bounds would be based on would never exceed the total number of distinct pre-test scores. This would result in wider bounds than necessary.

The confidence bounds are interpreted as follows: If the experiment were repeated again and again with an assumed constant population, then 95 % of the times (for $\alpha = 0.05$), the entire regression line would lie within the confidence band. If the confidence bands of two regression lines do not overlap in a large part of the pre-test score range, it is highly unlikely that the data stem from the same population. Such a result is strong evidence for an effect of the intervention in the experiment group.

16.3 EFFECT SIZE

The ongoing practice of reporting statistical significance has been heavily criticized for decades (e. g. Cohen, 1994; Johnson, 1999; Coe, 2002). Effect size measures (and confidence bounds) are often proposed instead. Correctly interpreted, the p-value only provides information about the probability of collecting this or more extreme data from the assumed population. It does not report the size of an effect. Furthermore, statistical significance depends heavily on the sample size. Effect size, on the other hand, is a statistic that quantifies the effect of, for example, an intervention and is independent of sample size. It is generally defined as the relation between difference in group means and the common standard deviation (e. g. Cohen, 1988, p. 20). In case the standard deviation is not common to both groups, a pooled standard deviation should be calculated, or alternatively, the standard deviation of the group (A or B) with the greater variance may be chosen as a reference. It is proposed to use the latter approach here to be conservative. Applying this concept to the DLG results in the following pre-test score-dependent expression:

$$d_{A,B}(X) = \frac{\hat{Y}_A(X) - \hat{Y}_B(X)}{\max(\bar{\sigma}_A, \bar{\sigma}_B)} \tag{36}$$

where the calculation of $\bar{\sigma}$ is based on the sum of squared errors, i. e. the deviation of individual measurements from the regression line:

$$\bar{\sigma} = \sqrt{\frac{SSE}{N-2}} = \sqrt{\frac{\sum i \sum j \left(Y_{ij} - \hat{Y}_j\right)^2}{N-2}}, \tag{37}$$

where

- Y_{ij} is the i-th observation on the j-th distinct pre-test score,
- $\hat{Y}_j$ is the value estimated by the regression line at the j-th distinct pre-test score, and
- N is the total number of observations.

The result is again discrete, but it may be interpreted and reported as a continuous line. In case a single value is preferred, $d_{A,B}(50\,\%)$ shall be reported. The discriminative effect can be considered by also reporting the effect sizes for the low- and high-achieving groups, e. g. $d_{A,B}(30\,\%)$ and $d_{A,B}(70\,\%)$.

16.4 QUANTIFYING THE DEGREE OF LINEARITY

A high degree of linearity means that the difference in mean post-test scores for a fixed difference in pre-test scores is comparable anywhere on the pre-test range, i. e. the slope is constant. As shown above, assuming a linear model results in two characteristic parameters for each course which are easily interpretable. A more complex model might result in a better fit, but also makes the interpretation less intuitive. Therefore, assuming a linear model seems to be the best option.

For testing this assumption when applied to specific data, a visual inspection is often sufficient. In critical cases, the F-test for lack of fit can be applied. One requirement for the application of the lack-of-fit sum of squares (SSLF), which is used to calculate the F-statistic, is that there are replicates, i. e. multiple observations of post-test scores for at least one pre-test score. (Neter et al., 1985, p. 124) Replicates are very likely to occur in pre-/post-test settings, given that the ratio of number of participants to number of distinct pre-test scores is reasonably high. Furthermore, the following assumptions must be met: for each pre-test score, the corresponding observations of the post-test scores are independent, normally distributed and have the same variance (Neter et al., 1985, p. 123).

The sum of squares due to error (SSE) can then be partitioned into the sum of squares due to pure error (SSPE) plus the sum of squares due to lack of fit (SSLF). The pure error addresses the deviations of the individual observations from the means, i. e. the "spread" of the individual data (see Figure 33). The lack of fit addresses the deviations of the means from the estimated regression line (see Figure 34),

weighted by the number of observations of each pre-test score, which provides a good estimate for the degree of linearity of the data.

$$SSE = SSPE + SSLF \tag{38}$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_j\right)^2 + \sum_{j=1}^{k} n_j \left(\bar{Y}_j - \hat{Y}_j\right)^2, \tag{39}$$

where

- k is the total number of distinct pre-test scores,
- n_j is the number of observations corresponding to the j-th pre-test score,
- Y_{ij} is the i-th post-test score observation corresponding to the j-th pre-test score,
- $\bar{Y}_j$ are the means over all n_j post-test score observations corresponding to the j-th pre-test score (see Figure 34), and
- $\hat{Y}_j$ is the estimated value at the j-th pre-test score.

The null hypothesis H_0, i. e. that a linear model is adequate, can be then tested by comparing the test statistic F^* to a critical value given by the cumulative distribution function of the F-distribution at a chosen confidence level:

F^* is the ratio of the lack of fit mean square (MSLF) and the pure error mean square (MSPE), which are obtained by dividing SSLF and SSPE by their respective degrees of freedom.

$$F^* = \frac{MSLF}{MSPE} = \frac{\frac{SSLF}{k-2}}{\frac{SSPE}{N-k}} \tag{40}$$

$$F^* \leqslant F(1-\alpha; k-2, N-k) \rightarrow H_0, \tag{41}$$

(Neter et al., 1985, pp. 127-130)[2]. For a good fit, the lack of fit should be small compared to the pure error, i. e. the $\bar{Y}_j$ should be close to the $\hat{Y}_j$. Large lack of fit statistics compared to small pure error statistics lead to large F^*-values. If these are greater than the critical value given by the F-distribution, the null hypothesis must be rejected. If the null hypothesis cannot be rejected, one can conclude that a higher-order model does not result in a substantially better fit than the linear one. The assumption of a linear model is then justified.

2 Note that the α in Equation (41) has the same meaning as in Equation (30) but their values can be chosen independently.

16.5 DEMONSTRATION

The DLG will be demonstrated using three data sets labelled treatment 1, 2 and 3. One data set was artificially generated for illustrative purposes, while the other two sets stem from real test administrations (see below). Here, the latter serve only for demonstration of the DLG. The interpretation of these data sets in terms of RBALM effectiveness will be discussed below in Part III.

Following a visual inspection for linearity, the F-test for lack of fit is applied to one of the real data sets for illustration. The regression lines with confidence bounds are calculated for all data sets. Two examples are presented to illustrate how to compare treatments.

16.5.1 *Linearity: visual inspection and Lack-of-Fit test*

Checking the appropriateness of a linear model is illustrated using the data shown in Figure 34. Visual inspection suggests that the linear model is adequate, as the mean post-test scores lie close to the regression line and do not deviate in a pattern suggesting a different model. In the extremes of the pre-test score range, the deviation of the means from the regression line is stronger than in the middle score range, but the sample of measurements is sparse in these regions (see Figure 33). Consequently, the respective means can be expected to be less precise estimators of the true value, and they also have less influence on the regression line compared to means from large sample sizes. Therefore, a deviation in the extreme ends can be tolerated under these circumstances.

When in doubt, the F-test for lack of fit should be applied. In the demonstration data, the administered pre-test consists of 30 items which would in theory result in $k = 31$ possible pre-test scores, but the two extreme scores $X_1 = 0$ and $X_{31} = 30$ did not occur. With $k = 29$, $N = 828$, and $\alpha = 0.05$,

$$F^* = \frac{13.56}{18.74} = 0.72 < 1.50. \tag{42}$$

The null hypothesis is not rejected. Hence, linearity can be assumed for this data. This result supports the impression from the visual inspection. The linearity checks for the other two data sets will be omitted as there is no additional value for demonstration. In each of the following two sections, two data sets are compared based on the DLG.

16.5.2 *Treatment 1 vs. treatment 2*

Figure 35 shows the resulting regression lines for treatment 1 and treatment 2. The maximum possible scores on the tests are reflected by the ranges of the axes: 30 points on the pre-test and 27 points on

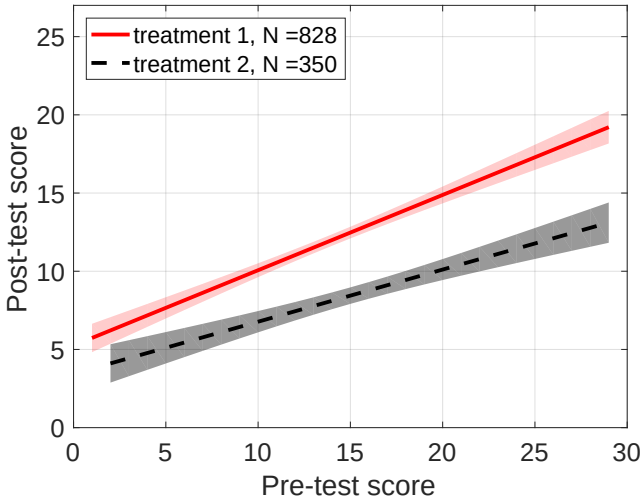


Figure 35: Comparing treatment 1 and treatment 2 with 95 % confidence bounds on the regression lines.

the post-test. The shaded areas represent the respective 95 % confidence bounds calculated by Equation (32). Compared to the variation in individual test scores (see Figure 33 and 34) the band seems quite narrow. This is because the band does not serve as a prediction for the next individual measurement, but instead as a prediction for the next regression line in case the experiment was repeated and a new sample of the same size was generated from the same population.

Treatment 1 results in a higher general post-instruction level of $\hat{Y}_{50\%,1} = 12.5$ compared to treatment 2, where $\hat{Y}_{50\%,2} = 8.4$. Furthermore, treatment 1 shows a greater discriminative effect with a slope of $b_{1,1} = 0.48$ compared to the slope of treatment 2, $b_{1,2} = 0.33$. The confidence bands in Figure 35 do not overlap except for the very low pre-test scores between 0 and 2 points where very few measurements exist. Neither estimated regression lines penetrates the other confidence band at any value on the pre-test score range. At a pre-test score of 50 %, there is a gap of 4.1 points between the means of the two treatments, 3.1 points between the inner confidence bounds, and 4.9 points between the outer confidence bounds. Because two uncertainties of α need to be accounted for, the total confidence is $(1-\alpha)^2$. For $(1-\alpha) = 0.95$, this results in a total confidence of 90.3 % that both regression lines lie within the confidence bands. If a larger total confidence is required, $1-\alpha$ must be chosen accordingly. The resulting interpretation would be that with 90.3 % confidence, a model student receiving treatment 1 with a pre-test score of 15 points will score between 3.1 and 4.9 points higher on the post-test compared to the same model student receiving treatment 2.

ANCOVA shows that the interaction term between treatment group and pre-test score is statistically significant at $p < 0.001$, i. e. the slopes

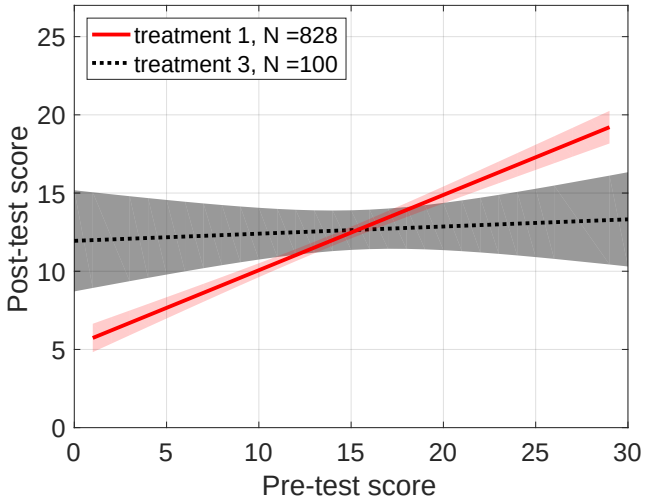


Figure 36: Comparing treatment 1 and treatment 3 with 95 % confidence bounds on the regression lines.

are different. Consulting Figure 35, one can conclude that the discriminative effect of treatment 1 is significantly higher than that of treatment 2. In other words, while treatment 1 results in superior expected post-test scores for all model students independent of pre-test score, students with higher pre-test scores tend to benefit even more from treatment 1 compared to students with lower pre-test scores. Finally, calculating the effect size based on Equations (36) and (37) yields quite large values:

$$d_{1,2}([30\%, 50\%, 70\%]) = [0.73, 0.94, 1.14]\,.$$

The established methods also show a difference between the treatments in terms of general post-instruction level, but the difference in discriminative effect is obscured by both, the scalar gain statistics such as g and the ANCOVA when falsely assuming a common slope (see Table 10). In case of g and c, the negative values for each data set can be misleading ($g_1 = -0.04$, $g_2 = -0.38$; $c_1 = -0.06$, $c_2 = -0.35$) as they suggest a loss of student expertise in the domain of interest.

16.5.3 *Treatment 1 vs. treatment 3*

Another example is presented in Figure 36 showing a comparison of treatment 1 to a third (ficticious) treatment. One feature to be illustrated here is a prominent difference in sample size, which shows in the different widths of the confidence bands. Another feature of this example are regression lines which cross within the pre-test score range. For the pre-test score range between $X_{13} = 12$ and $X_{20} = 19$, where the line of treatment 1 intersects the confidence band of treatment 3, a difference between the treatments cannot be detected. The

Table 10: Results from the comparative analysis of treatments using the different methods. All methods lead to the conclusion that treatment 1 is superior to treatment 2. Similarly, as long as only the first parameter of the DLG (the general post-instruction level) is considered, all methods lead to the conclusion that there is no difference between treatments 1 and 3. Inspection of the DLG's second parameter, reveals the strong difference in discriminative effect, suggesting that the most effective treatment is not the same one for students at high and low pre-test levels.

	Treatment 1 vs. 2	Treatment 1 vs. 3
g	$g_1 - g_2 = 0.34$	$g_1 - g_3 = 0.07$
c	$c_1 - c_2 = 0.29$	$c_1 - c_3 = 0.01$
ANCOVA	$\hat{Y}_{50\%,1} - \hat{Y}_{50\%,2} = 3.97$	$\hat{Y}_{50\%,1} - \hat{Y}_{50\%,3} = 0.00$
- common slope	0.43	0.42
DLG		
- general post-instruction level	$\hat{Y}_{50\%,1} - \hat{Y}_{50\%,2} = 4.1$	$\hat{Y}_{50\%,1} - \hat{Y}_{50\%,3} = 0.1$
- discriminative effect	$b_{1,1} - b_{1,2} = 0.15$	$b_{1,1} - b_{1,3} = 0.43$
- effect size	$d_{1,2} = [0.73, 0.94, 1.14]$	$d_{1,3} = [-0.55, -0.03, 0.48]$

general post-instruction levels hardly differ with $\hat{Y}_{50\%,1} = 12.5$ and $\hat{Y}_{50\%,1} = 12.6$, but the slopes indicate a strong difference in discriminative effect with $b_{1,1} = 0.48$ and $b_{1,3} = 0.05$. Treatment 3 seems to bring all students to the same level regardless of their pre-test score. The effect sizes are

$$d_{1,3}\left([30\%, 50\%, 70\%]\right) = [-0.55, -0.03, 0.48]\,,$$

indicating that stronger students benefit more from treatment 1 while weaker students benefit more from treatment 3. The established methods disagree: g shows a very small, and c and ANCOVA show no significant difference between the treatments (see Table 10). Again, the strong difference in discriminative effect is obscured by incorrectly assuming equal slopes in the ANCOVA or by ignoring the possibility of different effects for groups of different pre-test scores altogether when applying g or c.

In both examples, the general post-instruction level of the DLG aligned with the result from established methods, which support its validity. Additionally, the DLG's discriminative parameter discloses possible differences in the effect of instruction on groups with different pre-instruction levels.

17

DISCUSSION AND CONCLUSION OF THE ANALYSIS METHODS STUDY

Based on the argument that using pre- and post-tests that are non-identical is often a better option than using identical tests, several established methods for evaluating pre- and post-test data were presented. After discussing their shortcomings, the discriminative learning gain (DLG), based on a simple linear regression with confidence bounds, was introduced. It was shown how the line parameters can be interpreted and how inferences can be drawn about differences in learning success between courses using the same NIPPs.

Table 11 provides an overview illustrating the similarities and differences among the discussed methods, which are discussed in the following paragraphs. Note that although g and c are not intended for NIPPs, applying them to NIPPs is also considered for comparative purposes.

As mentioned above, g is no longer interpretable in a meaningful way if the post-test score is lower than the pre-test score. The apparent advantage of g and c, that courses can be compared regardless of the chosen test, is actually more a matter of IPPs or NIPPs than it is of the method, because the comparability requires the existence of a true neutral element. This "no learning" reference is generally unknown in case of NIPPs, but the zero in g or c is more readily misinterpreted as such than the line "post = pre" in case of the regression methods ANCOVA and DLG. This circumstance is a disadvantage of g and c in combination with NIPPs.

While g and c are scalar, the DLG-method returns two parameters for each line: the general post-instruction level and the discriminative effect. Comparing two data sets, the DLG hence returns four, while ANCOVA requires a common slope and therefore returns only three parameters. One disadvantage of a two-parameter learning gain is that it requires a larger sample size to produce reliable results, but in return, it offers a more fine-grained analysis in considering the discrimination effect.

Whereas g usually operates on class averages, c allows to consider statistical uncertainty, e. g. by calculating the standard error of the mean, although it is not always reported. The regression-based methods ANCOVA and DLG consider statistical uncertainties for instance by F-tests, confidence bounds, and effect size.

For g, the output itself is linear in the sense that a course with $g = 0.4$ achieved twice as much of what was still possible to learn compared to a course with $g = 0.2$. For c, linearity is only given sep-

Table 11: Illustration of similarities and differences among the methods. The plus and minus signs indicate whether the respective criterium in the first column is (+) or is not (-) fulfilled, and whether there is a high risk of misinterpretation (- -).

	g	c	ANCOVA	DLG
Allows for post < pre	-	+	+	+
Comparability across instruments				
- IPP:	+	+	+	+
- NIPP:	- -	- -	-	-
# of output parameters				
- one data set:	1	1	2	2
- two data sets:	2	2	3	4
Considers statistical uncertainty	-	+	+	+
Linear	+	+/-	-	-

arately on either side of zero. For ANCOVA and DLG, only the output *difference* between courses is linear (like a temperature scale).

All presented methods consider only a single covariate, the pre-test score. If more concomitant variables need to be included in the analysis, multiple linear regression (see e. g. Theobald and Freeman, 2014) might be more appropriate. Note that the multiple linear regression model assumes equal slopes like ANCOVA and that with multiple dimensions, the visualization in a single picture is lost.

Like all statistical tests and models, the DLG is also based on assumptions which result in requirements for the data that it can be used for. The model in Equation (22) assumes that all Y_i are independent normal random variables, with mean $E(Y_i) = \beta_0 + \beta_1 X_i$ and constant variance. Furthermore, the error terms ε_i are independent and normally distributed around zero (Neter et al., 1985, p. 49). Since the score ranges are finite, possible ceiling or floor effects always pose a threat to the normality assumption. If this assumption is violated, the DLG should always be interpreted with caution. The DLG furthermore requires the test pair to be the same for each course to be comparable. It is possible, however, to compare the *effect sizes* $d_{A,B}$ to $d_{C,D}$ even if the common test pair of courses A and B is different from the one used in courses C and D. Another prerequisite is that the DLG works on linked pre- and post-tests of individual students (see Direnga et al. (2016) for a proposal of how to link test scores). The linking procedure must be taken care of when administering the tests. In general, working with large data sets is favorable as larger sample sizes provide higher confidence. Judging by experience, the DLG can be used up-

wards of about 50 students. To achieve reasonably narrow confidence bounds, 150 or more students are desirable.

CAN THE DLG BE APPLIED TO IDENTICAL PRE- AND POST-TESTS?

The proposed method can also be applied to the special case of IPPs. Figure 37 shows an approximate distribution of pre- and post-test on data from IPPs[1] and the resulting regression line. In contrast to NIPPs, the line indicating "no learning" can be drawn. Consequently, with IPPs it is possible to make inferences about the course performance from a single regression line by comparing it to "no learning". However, if the effect to be measured is large, there will be a large difference between the pre- and the post-test scores. This makes it more likely for one of the two tests to have a floor or ceiling effect.

One consequence of a ceiling effect is a reduced estimated slope, especially when there is a non-negligible amount of observations in the higher pre-test score range. The estimated slope of the regression line will often be less than 1 because students in the high pre-test score range do not have many points to gain. The consequence is an underestimation of both, the discriminative effect as well as the general post-instruction level. In contrast to g and c which may be biased in favor of courses with higher pre-test scores, the DLG thus tends to be biased in the opposite direction when used on IPPs in the presence of a post-test ceiling effect. A pre-test floor effect would however produce a bias in the opposite direction, i. e. overestimating the discriminative effect, so that it could in principle reduce any bias introduced by a ceiling effect of the post-test. However, the best approach is to entirely avoid such effects by selecting appropriate instruments for the population.

Under consideration of all these aspects, the DLG is a valuable analysis tool applicable to NIPPs that provides two easily visualizable parameters to characterize the effect of a course or treatment in comparison to others. Especially when the differences in learning gains are small, confidence bounds on the regression lines help by giving insight how likely the lines, and therefore the effectiveness of treatments or courses differ. The discriminative effect as an additional output of the DLG can furthermore reveal differences in treatment effects on different subgroups that might be obscured by scalar learning gains or regression-based methods assuming equal slopes. Calculating an effect size at multiple pre-test score values allows to quantify this effect.

The DLG is applied in the following part of this dissertation to investigate the effect of research-based learning materials on student understanding of central course concepts.

1 FCI data from an introductory physics course at TUHH, $N = 495$.

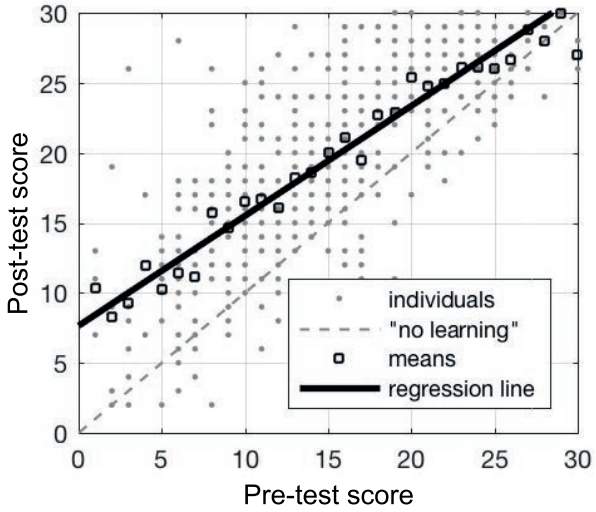


Figure 37: DLG applied to data from IPPs. Line can be compared to "no learning" reference, but a ceiling effect is more likely to occur.

Part III

THE EFFECTIVENESS OF TUTORIALS IN INTRODUCTORY ENGINEERING MECHANICS

"Careful assessment of student learning should be an integral part of the development of all printed and computer-based materials."

(McDermott, 2001, p. 1136)

18

INTRODUCTION TO THE INVESTIGATION OF RBALM EFFECTIVENESS

The previous parts were dedicated to answering research questions 2 and 3, which are concerned with the validity of measurements of student understanding (Part I) and how to analyze the measurement results (Part II). The answers found in those discussions form the basis for the investigation of the central research question, which is finally discussed in this part of the dissertation:

> RESEARCH QUESTION 1 Is instruction that uses research-based active-learning materials (RBALM) in the form of *Tutorials* more effective in promoting student conceptual understanding than traditional instruction in the context of introductory engineering mechanics courses?

The immediate effect of the Tutorials is assessed by means of pre- and post-test data, which were gathered in an introductory mechanics course over twelve consecutive years. In some of these years, the Tutorials were part of the curriculum, the other cohorts serve as a control group. Since instruction should enable students to develop sound understanding not only in the short run (e. g. until being tested in an examination) but also as a basis for further studies and subsequent professional practice, effectiveness in promoting understanding should not only be viewed as the immediate effect after instruction but should also consider long-term effects and the ability to transfer the understanding to different contexts. Therefore, the long-term effect of the materials is also investigated and discussed.

This part begins by describing the context of the investigation in Chapter 19: the student population of the course, the course content and learning objectives, changes made to the course over the years, and the implemented Tutorials. Chapter 20 describes the methods and data used to assess the immediate and long-term effects of the Tutorials. The analysis follows in Chapter 21. This part closes with a summary and conclusion in Chapter 22.

19 CONTEXT OF THE INVESTIGATION

The context of this study is the "Mechanics I" course at Hamburg University of Technology (TUHH). All study programs offered at this institution are set in the engineering disciplines. The investigated course has an enrollment of about 500 students or more as it serves several bachelor study programs, namely Mechanical Engineering (MB), General Engineering Science (AIW), Civil Engineering (BU), Naval Architecture (SB), and, newly introduced during the period of investigations, Mechatronics (MTB). Over the time span of this study, the course was subject to many changes, not only in terms of the implementation of Tutorials. The changes were motivated or necessitated by structural reorganizations or strategic decisions and will be illustrated in the following section. Subsequently, the course content and intended learning outcomes will be described.

19.1 COURSE STRUCTURE AND INSTRUCTIONAL FORMATS

Pre- and post-test data was collected over a time span of twelve years. Table 12 shows an overview of the course conditions for each of the twelve cohorts in terms of several variables, which will be described in the following paragraphs. As indicated in the first five columns of Table 12, Statics was initially taught during a first-semester course that included also Mechanics of Materials (MoM)[1]. The class time is documented in terms of three different course components: lectures (l) with little or no active participation by the students, plenary recitation sessions (p) in which students are presented with worked solutions to quantitative problems, and smaller group recitation sessions (g) where students actively solve problems under the supervision of a TA. In cohorts 1 to 4, approximately the first six out of 14 weeks were dedicated to Statics of rigid bodies (S)[2]. Out of the $4 + 1 + 2 = 7$ weekly units of class time, this corresponds to an "S-equivalent" class time of $1.8 + 0.4 + 0.9 = 3.1$ weekly units, where one unit corresponds to 45 minutes.

For cohorts 3 and 4, the entire course was re-scheduled to the second semester, primarily because the course was considered too challenging for the first semester. After two years, the course was divided into two separate courses with roughly one-half of the weekly hours, so that the heavy course load was distributed more evenly for the students. Subsequently, one course covered only S the other MoM.

1 German: Elastotatik
2 German: Stereostatik

For some study programs involved in this investigation, Physics is a co-requisite in some cohorts. This may be relevant as Physics instruction may have beneficial effects for the students' understanding of Statics and lead to better post-test results. Note that Physics as a *pre*requisite (as for cohorts 3 and 4) should play a minor role, as any beneficial effects would be accounted for already in the pre-test data.

The instructor changed several times. In the earlier cohorts, instructors A and B alternated, first annually, then biannually. Instructor C filled a gap before instructor D took over until the end of the data collection. Note that instructor A co-authored the instructional materials (Tutorials) under investigation.

In Table 12, N denotes the number of matched pre- and post-tests. It is often considerably lower than the total number of students enrolled in the course as attendance in lecture is not required, and pre- and post-testing took place during regular lecture time. While participation is strongly encouraged, it is neither formally required nor rewarded by grade incentives.

Instructional method (shown in the last column of Table 12) is the focus variable in this study, while all other observed changes are concomitant variables. "Traditional instruction" (Trad) refers to the instructional formats of lectures, plenary recitation sections, and group recitation sections, which were described above. In cohort 5, the Tutorials described in Section 4.3 were introduced. As an addition to traditional teaching, the implementation affected only the small group recitation sessions (g), while the lecture (l) remained unchanged. In the Tutorial cohorts (Tut), the plenary recitation (p) did not take place, instead, more time was allocated to the group recitation sessions (g). During about half of the time in each group recitation session, students solved quantitative problems (as in the earlier cohorts). During the other half, they worked on Tutorial worksheets.

In addition to the Tutorials, elements of *Just-in-Time-Teaching (JiTT)* (Novak et al., 1999) were introduced in cohorts 7 and 8. The students were given weekly reading assignments and online pre-tests. The results of the pre-tests were discussed at the beginning of each lecture, but no further changes were made to the course. In order to differentiate this "light version" from a full implementation of Just-in-Time-Teaching (JiTT), the abbreviation "JiTTl" will be used.

19.2 CONTENT, LEARNING OBJECTIVES, AND SELECTED TUTORIAL WORKSHEETS

The content of the investigated course has already been described in Section 11.1.2 as force systems and equilibrium, supports, trusses, weight force and center of mass, friction, ropes and chains, and internal forces and moments on a beam. The expected learning outcomes are documented in the TUHH course catalogue (Hamburg University

Table 12: Overview of changes in the course over a span of twelve years/cohorts. The first cluster (cohorts 1-4) are the early traditional cohorts (EarlyTrad), the middle cluster (cohorts 5-9) are the Tutorial cohorts (Tut), and the last cluster (cohorts 10-12) are the late traditional cohorts (LateTrad). The lecture time $t_l + t_p + t_g$ can be read as $t_l \times 45$ minutes lecture + $t_p \times 45$ minutes plenary recitation + $t_g \times 45$ minutes group recitation per week. Cohorts 2 and 9 are omitted from the analysis.
* no AIW-students sampled

Cohort	Semester	Content	Class Time	S-equiv. Class Time	Physics Co-requisite	Instructor	matched N	Instructional Method
1	first	S+MoM	4+1+2	1.8+0.4+0.9	MB, AIW, BU	A	134	Trad
~~2~~	~~first~~	~~S+MoM~~	~~4+1+2~~	~~1.8+0.4+0.9~~	~~MB, AIW, BU~~	~~B~~	~~61~~	~~Trad~~
3	second	S+MoM	4+1+2	1.8+0.4+0.9	-	A	120	Trad
4	second	S+MoM	4+1+2	1.8+0.4+0.9	-	B	223	Trad
5	first	S	2+0+2	-	AIW, BU	B	304	TutOnly
6	first	S	2+0+2	-	AIW, BU	B	367	TutOnly
7	first	S	2+0+2	-	AIW, BU	A	464	TutJiTTl
8	first	S	2+0+2	-	AIW, BU	A	361	TutJiTTl
~~9~~	~~first~~	~~S~~	~~2+(1)+2~~	~~-~~	~~AIW, BU~~	~~C~~	~~171~~	~~?~~
10	first	S	2+1+2	-	AIW	D	247	Trad
11	first	S	2+1+2	-	AIW	D	321	Trad
12	first	S	2+1+2	-	AIW*	D	217	Trad

of Technology, 2019), organized according to the German Qualifications Framework (DQR), which consists of four categories. Besides the knowledge and capabilities (or skills) categories, it includes personal competencies, which are divided into the categories social competencies and autonomy (for details see Bundesministerium für Bildung und Forschung (BMBF) and Kultusministerkonferenz (KMK) (2013)).

According to the current learning outcomes of the Statics course, the students completing the course are able to...

Theoretical Knowledge	... "describe the axiomatic procedure used in mechanical contexts;
	... explain important steps in model design;
	... present technical knowledge in stereostatics."
Capabilities	... "explain the important elements of mathematical/mechanical analysis and model formation, and apply it to the context of their own problems;
	... apply basic statical methods to engineering problems;
	... estimate the reach and boundaries of statical methods and extend them to be applicable to wider problem sets."
Social competence	... "work in groups and support each other to overcome difficulties."
Autonomy	... "determin[e] their own strengths and weaknesses and to organize their time and learning based on those."

(Hamburg University of Technology, 2019)

When comparing the content and the learning goals to introductory mechanics courses at other German universities, the biggest difference is that some universities also include MoM in their introductory mechanics course (as it was also done at TUHH during the first observations). Other differences are of little significance, for example, a slightly different emphasis on the different topics or a variation of the selection of individual topics, such that overall, the investigated course is comparable to other introductory mechanics courses at German universities.

The Tutorial worksheets selected for instruction were the same each year, only their order was modified occasionally. Changes to individ-

ual worksheets as a reaction to instructor or TA feedback were possible during the course of this study. In accordance with the course content described above, Tutorials addressing the following topics[3] were implemented:

1. Forces*
2. Forces and moments
3. Force couples
4. Equivalence of force systems
5. Supports and systems
6. Static and kinematic determinacy
7. Trusses
8. Friction
9. Ropes*
10. Internal forces and moments - discrete loads
11. Internal forces and moments - distributed loads

The worksheets marked by an asterisk(*) were adapted from the *Tutorials in Introductory Physics* (McDermott and Shaffer, 1998), the others were developed specifically for this course of engineering mechanics. The collection of the adapted and the newly designed Tutorials is published in Kautz et al. (2018).

The sessions were led by student TAs, who received a training of approx. 1 h per worksheet in which the authors of the materials were involved. This weekly preparation session consisted of taking the student perspective by working through the next Tutorial, resolving questions, and discussing the intended pitfalls. Furthermore, there was room for reflection on the previous Tutorial session where the TAs could give feedback on the worksheets.

3 Note that the English Tutorial names are stated here but the language of the implemented Tutorials was always German.

20 METHODS AND DATA

To assess the effectiveness of the Tutorials, the results of two studies are considered: the main study and the longitudinal study.

The main study was designed as a pre-/post-test study with NIPPs: the CATS as post-test and the FCI as pre-test, a pairing which is recommended by CATS developer Paul Steif (Atadero et al., 2015). In addition to investigating student understanding immediately at the end of the course, long-term effects are measured in the longitudinal study by administering the CATS again at a later time to the same students. For this purpose and under these circumstances, an IPPs study design is appropriate.

20.1 THE MAIN STUDY

The test administration of the CATS has already been described in Section 10.2. The FCI was usually administered during the first lecture in the same manner with a time limit of 30 minutes. In the last two cohorts, the instructor requested that the FCI was given not during lecture, but in the group recitation sessions by the TAs who were carefully instructed.

The analysis of the FCI pre-test / CATS post-test data is performed using the DLG introduced in Part II. Through multiple comparisons of cohorts or clusters of cohorts the effect of the Tutorials is estimated and separated from any effect of the concomitant variables. The clustering is described below. As indicated in Table 12, two cohorts (2 and 9) are omitted from the analysis. These decisions are justified as follows:

Cohort 2: The very small sample size in cohort 2 in comparison to the other cohorts ($N_2 = 61$) was very likely caused by the way the post-test was announced by the instructor, which seemed to rather discourage student participation. Consequently, cohort 2 is assumed to be biased more strongly by self-selection than the other cohorts and the data is omitted.

Cohort 9: For cohort 9, the instructor and two teaching assistants reported that the Tutorials were used, but not as consistently as intended. The instructor furthermore offered a plenary recitation section which targeted at-risk students and which was not in the official study plan. This makes cohort 9 difficult to categorize, both in terms of teaching and lecture time. Therefore, cohort 9 is omitted as well.

Table 13: Overview of cohort clustering for data analysis

Cluster	Cohorts	N
EarlyTrad:	the early traditional cohorts 1, 3 and 4	477
Tut:	the Tutorial cohorts 5, 6, 7, and 8	1496
- *TutOnly*:	the Tutorial cohorts 5 and 6 without any elements of JiTT	671
- *TutJiTTl*:	the Tutorial cohorts 7 and 8 with elements of JiTT	825
LateTrad:	the late traditional cohorts 10, 11, and 12	785

Cohort 12 is explicitly *included* in the analysis even though no data was collected for AIW-students due to scheduling issues. A comparison with the previous cohorts 10 and 11, which are comparable to cohort 12 in all recorded variables, revealed no indication for a significant difference in performance.

Education researchers do not often get the chance to investigate both, the change from a status quo to a teaching innovation, as well as the return to the previous setting, especially if the teaching innovation was found to be effective. The circumstances described in Section 19.1 allow a three-way comparison of the early traditional cohorts, the late traditional cohorts and the Tutorial cohorts to investigate temporal effects with two control cohort clusters, one chronologically before and one after the cohorts using Tutorials. For this purpose and the subsequent investigation of possible latent variable effects, the data is grouped in cohort clusters *EarlyTrad*, *LateTrad*, *Tut*, *TutOnly*, and *TutJiTTl* as indicated in Table 13.

The same circumstances allowed to investigate the level of retention of student conceptual understanding in a longitudinal study, and especially to compare this retention based on instructional method. For this purpose, the CATS was administered once more in a different setting as a retention test, the so-called *reTest*, which is described in the following section.

20.2 THE LONGITUDINAL STUDY

In contrast to the continuous collection of post-test data, the reTest data were obtained at two distinct events in November 2014 and 2017. The target population were all students whose post-test data were in the database up to the respective reTest event. As a result, post-test/reTest score pairs with various retention intervals (RIs), i. e. the time interval between post-test and reTest, were obtained. The range of possible RIs spans one to twelve years.

While the post-tests could be administered in one single class during lecture time, this was not possible for the reTest as the aim was to

reach students from all cohorts. Instead, the events were advertised on campus (and in the first run also to the alumni network, which proved to be inefficient). The advertisement channels used were mailing lists to all students and research assistants, posters and flyers, as well as short announcements in selected lectures. To acquire as many participants as possible, the test could be taken in paper format on campus or online. To reduce the self-selection effect, material incentives[1] were offered. Before starting the CATS, reTest participants were asked to fill out a questionnaire asking for demographic data on their physics and mechanics background, on their personal teaching activity, as well as on their study progress. Of these aspects, only personal teaching activity is included in the analysis in this dissertation.

Using only the reTest 1 data, it is impossible to attribute any observed effect to either length of RI or type of instructional method because they are coupled see (see Direnga et al., 2015a). With the instructional method changing back to Trad, a second reTest was sufficient to uncouple RI and instructional method. Data for RIs of two to three years as well as six to eight years for both types of instructional method is available, but as many students graduate after 6 years, the number of participants for the longer RIs is very small, so that only the RIs 2, 3, and 6 are selected for analysis (see Figure 38).

The data is analyzed separately for subgroups differing in the three variables RI, instructional method, and personal experience as TA. To account for the difference in post-instruction CATS scores among the groups with different pedagogies, the normalized change metric according to Equation (16) is applied to the observed change from posttest to reTest performance. In addition, absolute scores are compared for the different subgroups.

1 A lottery for all participants with four drones and four audio speakers as prizes, as well as chocolate for the offline participants only

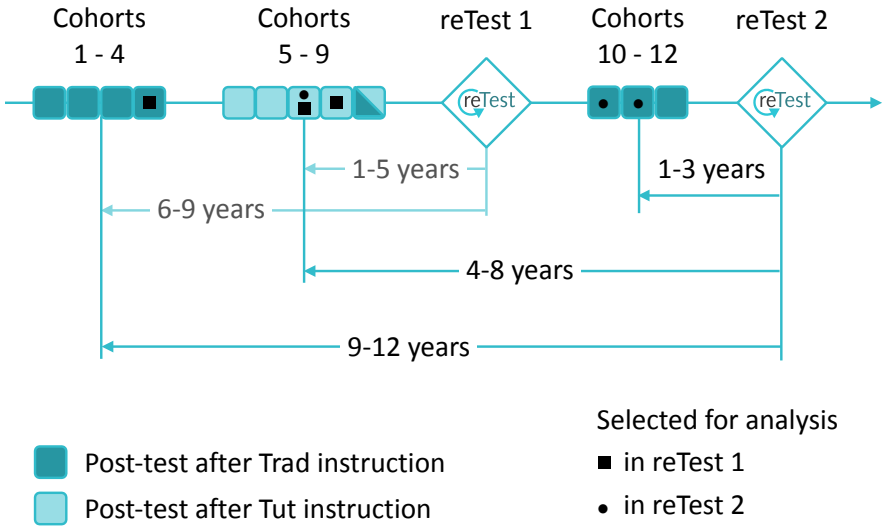


Figure 38: Study design timeline

21 ANALYSIS

In this chapter, the results of the investigation regarding the effectiveness of Tutorials in introductory engineering mechanics are presented. First, the three data sets of the *EarlyTrad*, *Tut*, and *LateTrad* cohorts are compared. To strengthen the evidence that the difference revealed by this comparison can be attributed to the Tutorials, a possible contribution by the latent variables that changed over the years is investigated as a second step. Finally, the reTest sample and results are described and the long-term effects of the Tutorials are discussed.

21.1 ANALYSIS OF THE MAIN STUDY

21.1.1 *Comparing Tutorials to traditional instruction*

Figure 39 shows the resulting regression lines and confidence bands when applying the DLG to the cohorts using Tutorials and the cohorts not using Tutorials (*EarlyTrad* and *LateTrad*). Note that the linearity assumption was successfully checked for all DLG lines discussed in this chapter, both visually and by the F-test for lack of fit. The large sample sizes lead to high confidence in the models, indicated by the narrow 95 % confidence bands.

Two aspects immediately come to attention: (1) there is no significant difference between the regression lines representing the Traditional cohorts, even though many other variables varied, such as instructor, course content, and lecture time. The performance of a Traditional course seems to be time-invariant, and any effects of the concomitant variables seem to be very small (or cancel out). (2) The regression line of the Tutorial cohorts clearly lies above the (nearly identical) lines of the Traditional cohorts. It does not penetrate the other confidence bands at any pre-test score level, which leads to the conclusion that there is a statistically significant effect on the entire range. The pronounced gap between the lines indicates that the effect is not only significant, but also prominent in size. Table 15 shows the effect sizes at three different pre-test score levels for each pairwise comparison of the lines. Comparing the Tutorial cluster to either of the traditional clusters reveals substantially larger effect sizes (up to 0.83) than comparing the two traditional clusters to each other (at most 0.09). If the latter result is interpreted as the impact of a variable called *time*, which may include changes in other unobserved variables, the effect size of this variable is indeed negligible.

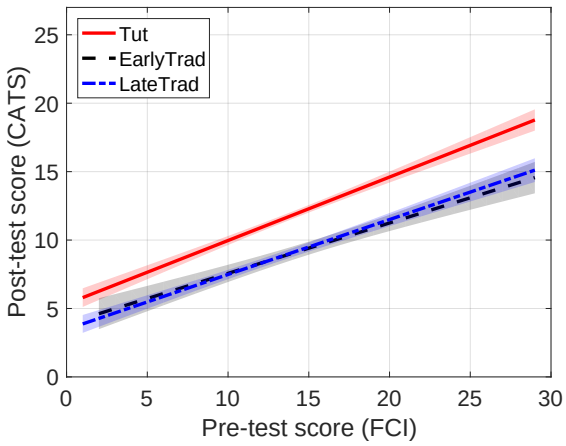


Figure 39: Going back to traditional teaching shows in the data: while the cohorts using Tutorials stand out, the recent Traditional cohorts' DLG line (LateTrad) resembles the one from before the introduction of Tutorials (EarlyTrad).

21.1.2 *Ruling out a possible effect of other variables*

The introduction of Tutorials was not the only change to the course or the context. It is therefore justified to ask to which extent other variables might cause the effect. In the following, the possible influence of the observed latent variables *JiTTl*, *lecture time*, *physics*, *instructor*, as well as gender composition of a cohort is discussed.

JITTL If the additional elements of JiTT were causing the effect instead of the Tutorials, the *TutOnly* cohort clusters would be comparable in their performance to the *Trad* cohorts. Consequently, the *TutJiTTl* DLG line would have to lie even further above the *Tut* DLG line in order to result in the large *Tut*-vs.-*Trad* effect shown in Figure 39. To test this hypothesis, we compare the *TutOnly* cohorts to the *TutJiTTl* cohorts (see Figure 40). As the difference, if any, is very small (effect size at most 0.2), it is justified to attribute the major part of the effect to the Tutorials instead of JiTTl.

This does not mean that JiTT as promoted by Novak et al. (1999) is ineffective. The data does not allow to investigate the effect of JiTTl without Tutorials. It is thus possible that JiTTl alone could have caused a similarly large effect as seen by the Tutorials alone, although it is more likely that the merely negligible effect of introducing "JiTT light" on top of the Tutorials is grounded in the partial implementation of JiTT, as "[...] it is consistently recognized that 'partial implementation' of complete educational reform leads to limited or no improvement in student conceptual mastery" (Finkelstein and Pollock, 2005).

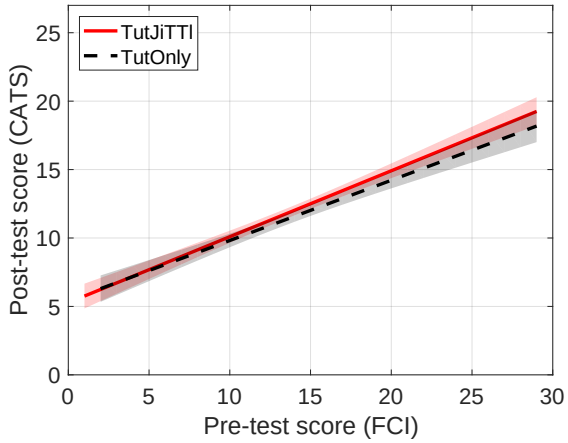


Figure 40: Investigating the effect of JiTTl. The aggregated cohorts 7 and 8 (with JiTTl, instructor A) are compared to the aggregated cohorts 5 and 6 (without JiTTl, instructor B). A significant difference that could be attributed to the use of JiTTl cannot be seen.

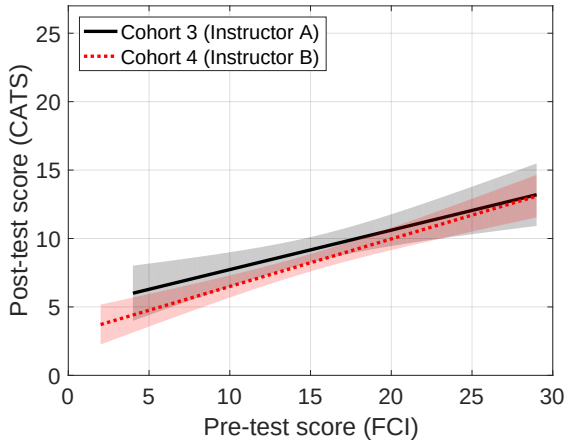


Figure 41: Investigating the effect of different instructors. The effect, if any, is small for low scoring students and negligible to non-existent on the rest of the pre-test score range.

CLASS TIME Could the observed effect be explained by the difference in class time? Literature suggests that this is not the case, although most studies on the effect of additional class time are set in the school context: For example, Kidron and Lindsay (2014) report in their meta study that if any effects were found, they would be small. Joyner and Molina (2012) point out that "the amount of instructional time is not so important as how that time is spent." Even specifically in the Statics context in higher education, Burkhardt (2015), who investigated the exam performance of previously identified at-risk students, found that one additional hour on an otherwise three-hour Statics course (33 % more class time) has no statistically significant effect.

The S-equivalent class time in the traditional cohorts 1 to 4 is lower by 0.9 SWS (23% less class time) than in the Tutorial cohorts, resulting in a non-conservative comparison, but the additional time was spent on the Tutorials, which is not just "more of the same". Considering the results from literature, the size of the observed effect can only be explained by the Tutorials, while it cannot be explained based only on the extra class time. It can be concluded that the Tutorials were effective and that they were given the appropriate amount of time to be implemented as intended.

The effect of the plenary recitation session that was re-introduced for cohorts 10 to 12 along with the return to the traditional setting cannot be quantified, but it allows a conservative interpretation of the results. Based on the reasonable assumption that the additional p-session, which provides 20 % more class time, favored the control group, the true effect of the Tutorials could thus be even larger than observed.

PHYSICS Additional physics instruction may introduce a bias. Splitting cohort clusters into the study programs with and without physics reveals no coherent picture. In the *LateTrad* cluster, there is no difference at all while in the *Tut* cluster, the study programs *without* physics tend to perform slightly better (effect sizes -0.18, -0.22, -0.26). This trend is opposite to the one expected but as the effect (if any) is quite small, the question whether or not a physics course parallel to the investigated mechanics course introduces a significant bias can be negated.

INSTRUCTOR A strong effect of the instructor is not expected as it has been observed that "[s]tudents in equivalent physics courses with different instructors are remarkably similar in the way they respond to certain kinds of questions, [...]" (McDermott, 2001). There is no reason to doubt that this observation is transferable to engineering courses, and it is supported by the data: If one assumes nonetheless that the effect is not caused by the Tutorials but by the different in-

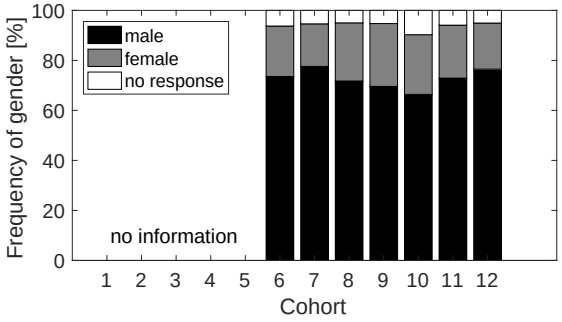


Figure 42: Gender composition remains nearly constant over the years. No gender bias to be expected in the results.

structors, one would be led to the conclusion that instructors A and B must in general achieve higher gains than instructor D. Referring again to Figure 39, it is evident that this is not true. Cohorts 1 to 4 were taught by instructors A and B, cohorts 10 to 12 by instructor D. As stated before, the DLG lines are nearly identical with high confidence. Only if class time is considered, one might argue that instructors A and B achieved the same DLG line with 1.9 SWS (= 38 %) less class time. Although this is a substantial amount, the additional time was spent on "more of the same" activities. Therefore, we conclude that instructors A and B cannot in general achieve higher gains than instructor D.

The data allow for another way to look at a possible effect of the instructor. When comparing cohort 3 to cohort 4, which only differ in the variable *instructor*, the DLGs are very similar (see Figure 41) and the effect sizes are small. Considering all pieces of evidence presented above, the assumption that the large effect observed in Figure 39 is caused by the instructor cannot be supported by the data.

GENDER BIAS The data regarding gender was only collected for cohorts 6 and later. The only available options were "male" and "female", some students chose to give no response. The composition did not change noticeably over the years (see Figure 42). Therefore, any possible gender bias on the concept inventories has no impact on the interpretation of the results with respect to the effectiveness of the instructional material.

21.1.3 *Effect sizes*

As noted in Section 16.3, the pre-test score dependent effect size can also be represented as a line. The three pre-test score levels of 30, 50 and 70 % at which the effect sizes are presented in Table 15 were chosen to present the effect size as a numerical value and to illustrate its dependency on the pre-test score. In addition, Figure 43 shows a

Plotted in	Cluster (Cohorts)	DLG parameters b_0	b_1
Figure 39	Tut (5, 6, 7, 8)	12.3	0.47
	TradBefore (1, 3, 4)	9.3	0.37
	TradAfter (10, 11, 12)	9.5	0.40
Figure 40	TutJiTTl (7, 8)	12.5	0.48
	TutOnly (5, 6)	12.0	0.44
Figure 41	Instructor A (3)	9.0	0.30
	Instructor B (4)	8.2	0.34
Figure 47	Tut (5, 6, 7, 8)	12.3	0.47
	Trad (1, 3, 4, 10, 11, 12)	9.4	0.39

Table 14: Overview of all the DLG parameters. It is evident that the Tutorial cohorts have a much higher general post-instruction level (b_0) than the Traditional cohorts.

graphical representation of the effect size for all pre-test score levels, which highlights the strong dominance in effect size associated with the Tutorials over other variables, i. e. *time* (and the latent variables associated with it), *JiTTl*, and *instructor*. Care must be taken not to confuse this representation with the DLG regression lines.

Plotted in	Compared clusters			Effect size at pre-test score level 30 %	50 %	70 %
Figure 39	LateTrad	vs.	EarlyTrad	-0.0	0.0	0.1
	Tut	vs.	EarlyTrad	0.5	0.7	0.8
	Tut	vs.	LateTrad	0.6	0.7	0.7
Figure 40	TutJiTTl	vs.	TutOnly	0.1	0.1	0.2
Figure 41	InstrA	vs.	InstrB	0.3	0.2	0.1
Figure 47	Tut	vs.	Trad	0.6	0.7	0.8

Table 15: Effect sizes

21.2 ANALYSIS OF THE LONGITUDINAL STUDY

21.2.1 *The reTest sample*

In total, 656 participations were counted in both reTest events, however, the majority of the data was not considered in the analysis for several reasons:

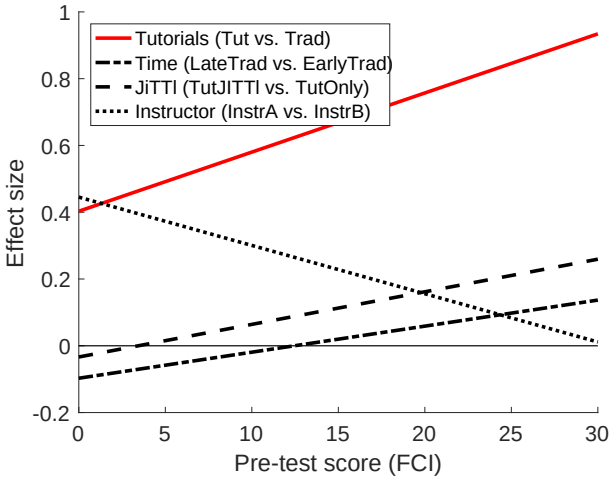


Figure 43: Effect sizes in dependence of pre-test score.

First of all, it was expected to observe cases of unserious participation, which is defined as less than 9 items answered, or less than 10 minutes time spent on the test (as devised by Steif and Hansen, 2007), provided the total score does not exceed 14 out of 27 points. The total number of cases of unserious participation was 83, which were all observed among the online participants, possibly due to unsupervised and distraction-rich test-taking settings. Furthermore, 24 cases had to be eliminated because they could not be uniquely matched to a post-test. The result is a total number of 549 valid participations over all RIs.

In addition, certain RIs were omitted. The small sample sizes n_i in the subgroups for RIs 7 and 8 ($n_7 = 3, n_8 = 1$ for *Trad* and $n_7 = 7, n_8 = 2$ for *Tut*) limit the analysis to RIs 1 to 6. Furthermore, RIs 1, 4, and 5 are unsuitable for a comparison of the influence of instructional method as each only provides data for one of the methods (see Figure 38). Omission of RIs 7 and 8 as well as 1, 4, and 5 reduces the sample size of the selected subset to a total number of 282 participations.

Among those 282, 98 were categorized as having TA-experience in relevant subjects which include Mechanics, Physics, Engineering Design as well as others specified by the participants. The majority, i. e. 178 of the 282 participants (63 %), experienced instruction using Tutorials.

In general, a representative sample is required in order to make inferences from the sample to the population. As this sample is self-selected, it does not necessarily represent the population. Especially TAs seem to be overrepresented and will thus be analyzed separately. Comparing the distributions of the post-test scores from the sample to the ones from the population reveals that the stronger students are slightly overrepresented in the sample. In Figure 44, the effect of

instructional method on the representation is investigated. The relative frequencies of the possible post-test scores in the sample are compared to the ones in the population for the selected RIs. Transformation to the log2 scale allows for easy visualization of over- and underrepresented scores. For the *Tut*-participants, the higher scores are overrepresented while the lower scores are underrepresented. This tendency is far less pronounced for the *Trad*-groups. Still, this bias may limit the generalizability of the interpretation of the results in the sense that it may not be valid for the group of low performing students. The reason for the bias is unclear but there are likely two main factors: (1) Self-selection bias can play a role as the reTest situation resembled an "opt-in" (but with material incentives) instead of an "opt-out" as in case of the in-class pre-/ and post-tests. The material incentives may have reduced the size of this bias. (2) The advertisement for the reTest events did not reach (supposedly weaker) dropout students. These students do not belong to the population of interest such that a certain underrepresentation of low-scoring participants is expected.

Comparing the post-test scores among the different subgroups (Figure 46) reveals significantly higher average scores for *Tut* over *Trad* (as seen in the DLG analysis above), slightly higher average scores for TAs over non-TAs, as well as overall slightly higher average scores in RI 2 over RIs 3 and 6.

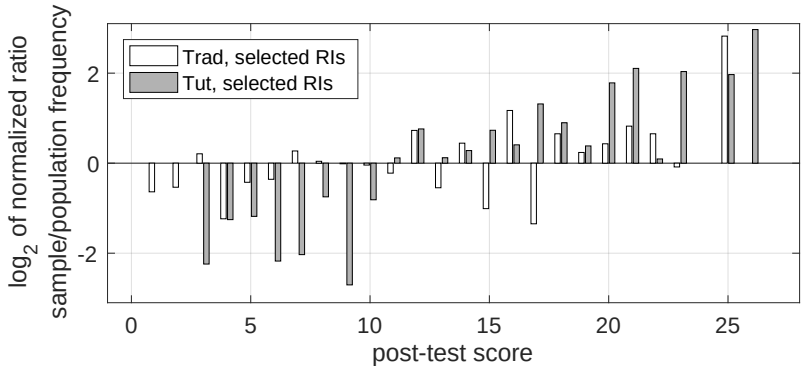


Figure 44: Does the sample represent the population for each pedagogy subgroup? A value close to zero means that the sample represents the population well in the respective posttest score value. A value of positive/negative y indicates an overrepresentation/underrepresentation by the factor 2^y. Note that the cases for which either frequency is zero would have to be represented by $\pm\infty$ and are therefore not displayed.

21.2.2 *The reTest results*

The results are first examined in terms of normalized change before examining the absolute scores achieved by the various subgroups

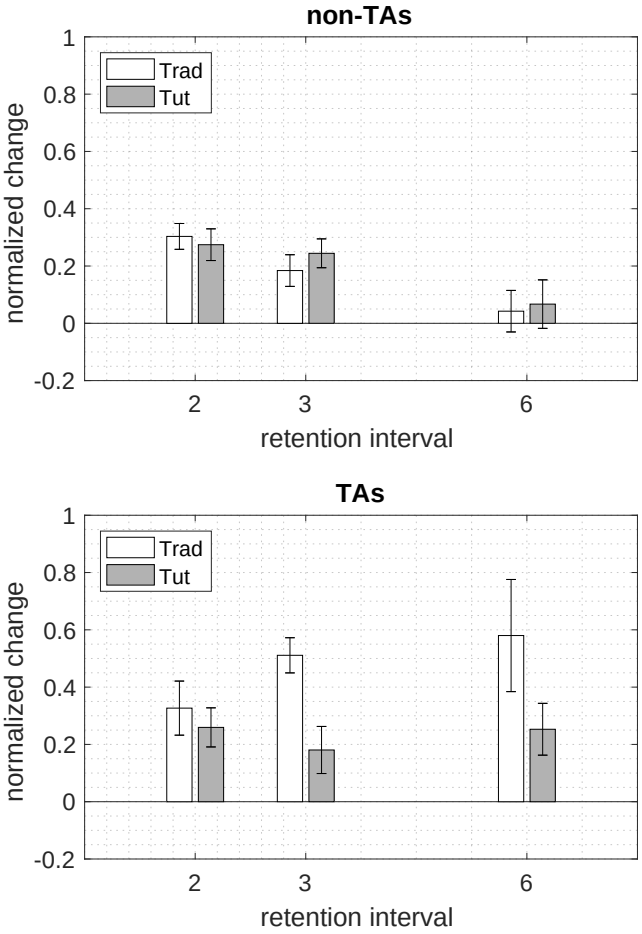


Figure 45: Mean normalized change for the various subgroups. The error bars indicate the standard errors of the means.

which differ in terms of the three variables pedagogy, personal TA-activity, and RI.

Figure 45 shows the average normalized change for the various subgroups. All subgroups exhibit on average positive normalized changes, interpretable as a further gain in understanding after formal instruction on the content had been completed. For students with TA-experience (subsequently also referred to as TAs), instructional method seems to be a relevant factor in combination with RI. The traditionally taught students with TA-experience show a much larger gain in RIs 3 and 6 than the ones that were taught with Tutorials. The gain seems to increase with time for the traditionally taught students with TA-experience while it remains constant for the interactively taught students with TA-experience. For the students without TA-experience (subsequently also referred to as the non-TAs) instructional method seems to be an irrelevant factor to the average normalized change judged by the overlapping error bars in every RI.

What does this mean in terms of absolute score? No difference in normalized change with respect to instructional method would mean that the instructional method group with higher average post-test scores also ends up having higher average reTest scores. Due to the measurement uncertainty (visualized by error bars), this effect is not always statistically significant. Figure 46 shows the mean post- and reTest scores. As seen by the normalized change, all subgroups exhibit a mean absolute gain from post- to reTest, indicated as the grey part of the bars. First, the results of the students without TA-experience are examined. While the *Trad* group with an RI of two years has managed to reach similar average reTest scores as their peers who learned with Tutorials, the difference between the instructional methods remains visible in the subgroups with RIs of three and six years, where the initial difference between the average post-test scores was larger. The average reTest scores among the students with TA-experience show no significant difference with respect to RI or instructional method. The TA-experience obviously helps *Trad* students to reach the same level of understanding as their colleagues who initially learned the concepts of statics with the help of Tutorials.

The large gain by TAs in the traditionally taught groups poses the most radical effect seen in the data. A possible interpretation is that personal TA-activity results in interactive engagement with the concepts, which leads to high gains. It can be assumed that many TAs carry out their teaching activity for more than one year. Therefore, TAs with higher RIs can be assumed to have repeatedly engaged in the concepts and gained even more understanding. Traditionally taught students thus achieve high gains by means of their TA-activity. Those students who had interactive instruction, already experienced these gains before the post-test and thus do not exhibit such high gains from post- to reTest.

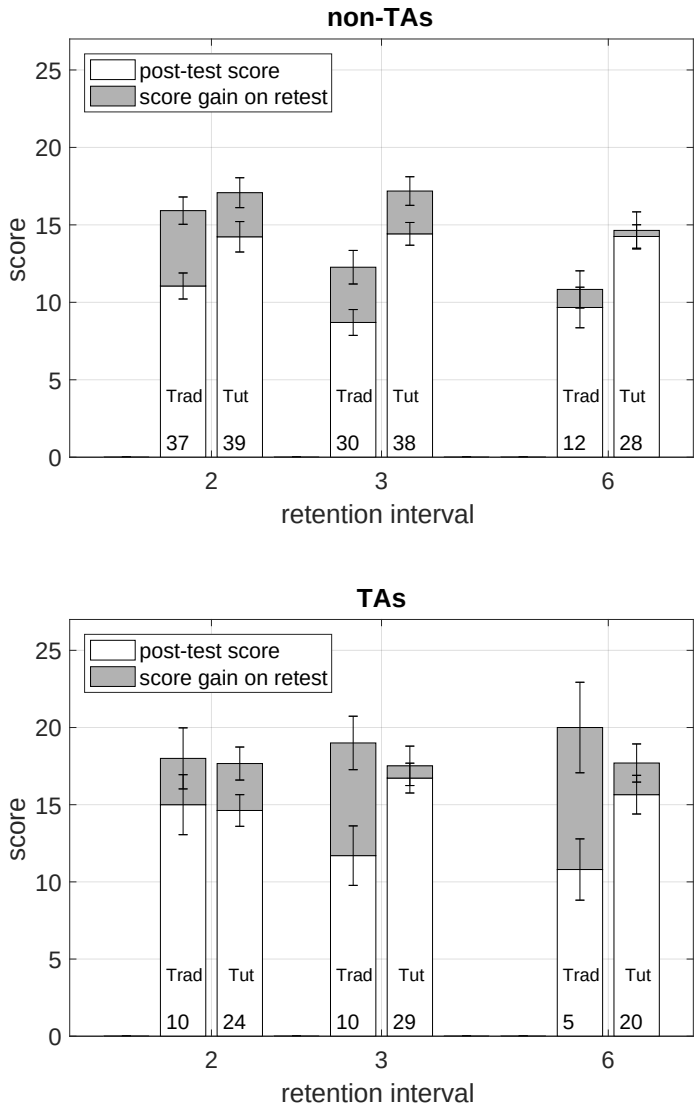


Figure 46: Mean CATS scores on the post- and reTest for the various subgroups. The reTest score is displayed as the score gain on the reTest stacked onto the post-test score. The error bars indicate the standard errors of the mean post- and reTest scores (not the error of the score gain). Subgroup sample sizes are displayed at the bottom of the bars.

22

SUMMARY AND CONCLUSION OF THE INVESTIGATION OF TUTORIAL EFFECTIVENESS

This part of the dissertation investigated the effectiveness of research-based active-learning materials (RBALM) in the form of *Tutorials* on the development of student conceptual understanding in an introductory mechanics course. Using the Discriminative Learning Gain (DLG), non-identical pre- and post-test data from 2774 students in 10 cohorts were analyzed. Furthermore, the long-term effects were investigated in a retention study.

A general limitation of pre-/post-test studies is that the level of understanding is dynamic and non-monotonically changing within a week or even smaller time scales (Sayre et al., 2012). The measurements are thus only snapshots of the individual student's understanding at two instants in time, which may have been much better or worse a week earlier or later. Because the measurements were always taken roughly at the same time and the analysis is based on cohort average scores, such an effect is not expected to affect the reliability of the results. The fact that the DLG line of the *LateTrad* cohorts reproduces the one of the *EarlyTrad* cohorts so well supports this view and allows to combine the two traditional clusters to compare only cohort clusters *Tut* to *Trad*, as summarized in Figure 47. This analysis shows that using Tutorials results in a higher average level of conceptual understanding for students of all pre-test score levels. The general post-instruction level differs by 2.9 ± 0.5 of 27 points ($(11 \pm 2)\,\%$). The lines do not intersect anywhere on the pre-test score range so that the *Tut* students have a higher expected post-instruction level than the *Trad* students, independent of pre-instruction level. Furthermore, the slopes differ by 0.08 ± 0.10, indicating a slightly higher discriminative effect when using Tutorials. The effect sizes according to Equation (36) for the 30 %, 50 %, and 70 % pre-test score levels are 0.6, 0.7, and 0.8, respectively.

Another general limitation of pre-/post-test studies is that effects from all (observed and unobserved) events between the two tests are captured, not only those induced by the course. A possible influence of other variables that changed in the period of data collection, such as instructor, class time etc., was examined and discussed. No other variable could be associated with such a strong effect as the Tutorials.

It should be emphasized that, against common prejudice, the Tutorials do *not* support only weak students. On the contrary, the stronger students tend to benefit even more from the Tutorials than the weaker ones do. In addition, the Tutorials may have had a positive influence

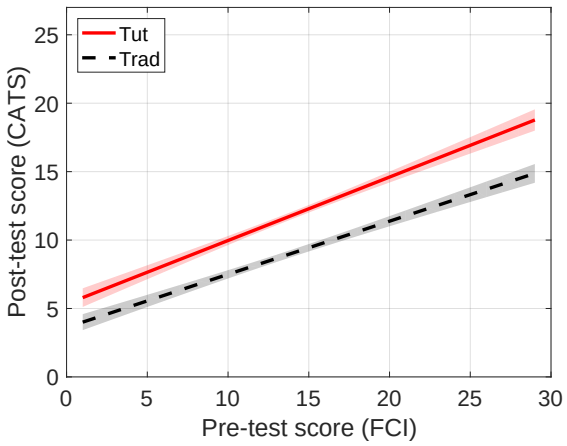


Figure 47: Comparing the DLGs of the cohorts using Tutorials and the cohorts not using Tutorials.

on attendance as can be concluded from the higher numbers of students taking part in the post-test in the *Tut* cohorts compared to the *Trad* cohorts. Although there is no attendance data available from the group recitations where the Tutorials were implemented, the larger sample size on the test hints at a higher attendance in lecture which also suggests higher attendance in the group recitation. This is a positive effect as the Tutorials can only help those students who come to class.

One may argue that it could have been primarily the strong students who chose to stay away from class in the *Trad* cohorts, so that the results appeared worse than in the *Tut* cohorts only due to sampling bias. While this case cannot be completely ruled out, there are two aspects that make it rather unlikely: First, strong students in terms of the post-test likely also score high on the pre-test. If all data points were missing in this region, the DLG could still extrapolate based on the data in the lower pre-test score region. If only the higher data points in both, pre- and post-test, were missing, the slope would be slightly underestimated, and the observed discriminative effects may actually be equal for both types of instruction. One would have to revise the conclusion that the stronger students benefit more than the weaker students, but the main conclusion remains valid: students of all pre-test scores tend to benefit from the Tutorials. Second, and maybe more importantly, the question then arises *why* the stronger students chose to stay away from class in the *Trad* cohorts, but not in the *Tut* cohorts. If that was indeed the case (which cannot be proven here), the strong students must have the feeling that instruction with Tutorials has value for them while it does not have value without the Tutorials.

Considering all the available evidence, it is concluded that using the Statics Tutorials complementary to traditional lecture and stan-

dard problem-solving exercises results in a higher average level of conceptual understanding for students of all pre-test score levels. As literature suggests that training of TAs in socratic dialogue is essential for the success of the Tutorials (e. g. Koenig et al., 2007), and the amount of TA-training was minimal, the materials were thus implemented in non-optimal but realistic conditions and still show a considerable effect, which may even underestimate the potential of the materials. This higher level of conceptual understanding is sustainable, as it could not only be observed at the end of the course but also up to 6 years later.

FINAL CONCLUSION

23 DISCUSSION AND CONCLUSIONS

This dissertation investigated concept inventory data administered as non-identical pre- and post-tests (NIPPs) in an introductory Statics course to assess the effectiveness of research-based active-learning materials (RBALM) in fostering student conceptual understanding. The instrument administered as post-test, the Concept Assessment Tool for Statics (CATS), was closely examined to ensure that the test results can be validly interpreted as a measure for student conceptual understanding in the given context of a German higher education introductory mechanics course (Part I, conclusion: Chapter 13). Amongst other aspects, the prior assumption that the CATS is not well interpretable by students before formal instruction on mechanics could be confirmed, which justifies and necessitates the use of a different, more suitable instrument as pre-test, i. e. the Force Concept Inventory (FCI). The challenge of interpreting data collected with NIPPs led to the development of the Discriminative Learning Gain (DLG), a two-parameter quantification of the difference in learning success between courses (Part II, conclusion: Chapter 17). Using this analysis method, FCI pre- and CATS post-test data from 10 cohorts, of which some used RBALM in the form of *Tutorials*, were compared. In addition, a longitudinal study was conducted to investigate the long-term effects of the Tutorials. (Part III, conclusion: Chapter 22).

23.1 USING CATS TO ASSESS COMPETENCIES IN STATICS IN A GERMAN CONTEXT

SUMMARY The study on the validity of interpreting results of the CATS when administered in the German higher education context allows to conclude that the total score at post-instruction level of individual students (and even more so of group averages as they are used to assess the effectiveness of the Tutorials) may be interpreted as a proxy for the level of conceptual understanding of statics.

The CATS was found to focus on a set of concepts that intersects largely with the one considered fundamental and essential by instructors and textbook authors. Apart from a few exceptions, the items were found to be interpretable by the students such that incorrect answers can be attributed to lack of statics understanding. The population was shown to be similar to the US students in their preference for certain distractors. Statistical analyses of the test data show that the overall performance of the CATS as a measurement instrument is good in the German context, although slightly less so when compared

to the US data, which may in part be caused by the (un-)familiarity with the representation of the systems on the items. Measurement precision may improve by allowing students more time on the test. For further suggestions for improving the CATS see Appendix G.

WHY WAS THE FCI NOT REVALIDATED? In contrast to the CATS, the FCI has not been inspected for validity in this dissertation. Even if the FCI acts only as pre-test, which is required to filter out possible differences in pre-instruction understanding of the Newtonian force concept but which otherwise plays only a subordinate role, correct interpretation of the DLG requires it to measure precisely. Unlike the CATS, the FCI has been widely used in Germany before without apparent problems with respect to interpretation of the total score (e. g. Girwidz et al., 2003; Schecker and Gerdes, 1999). Only doubts in its diagnostic power of student misconceptions in the sub-scales have been expressed for instance by Schecker and Gerdes (1999)[1]. As this dissertation does not make use of the sub-scales but relies only on the total score, it is justified to assume the FCI to be a valid measure for understanding of the Newtonian force concept.

IS THE CATS VALID AS A RETEST? By using the CATS as a retention test, as described in Part III, it was implicitly assumed that the interpretation of the results are valid when the test is administered to students in higher semesters. This assumption had to be made for practical reasons because another revalidation for the various subgroups goes beyond the scope of this dissertation. It is, however, a reasonable assumption to make because it is very unlikely that the population changes dramatically in aspects relevant to validity. First of all, the students are adults, not children. The latter would experience a much greater development in the investigated time frame and validity should indeed be questioned. Second, student experiences between reTest and post-test mainly consist of courses on entirely different content, which is different from the experience made from pre- to post-instruction. Instruction *on the content of the instrument* is much more critical for validity than any other instruction.

It is furthermore assumed that no memory effects from the post-test are influencing the results of the reTest. This assumption is reasonable as experience from studies using identical pre- and post-tests shows that these memory effects do not even occur when the interval between test administrations is much shorter, i. e. the length of one course (Henderson, 2002).

1 in the context of the German secondary education, and relating to a previous FCI version

23.2 ANALYZING CI DATA IN A NIPP STUDY DESIGN

SUMMARY The DLG was developed in this dissertation to address specifically the issue of analyzing NIPPs data. In comparing the DLG to other established analysis methods, it was demonstrated that the DLG poses a well-interpretable, two-parameter quantification of differences in course performance that is more fine-grained than scalar measures, while remaining easily visualizable unlike multiple linear regression. Limitations of the method involve requirement of a linear relationship as well as a link between student pre- and post-test data.

WHICH TESTS CAN BE USED WITH THE DLG? The DLG is most robust if the pre-test data has a reasonable spread. Floor and ceiling effects should be avoided by choosing instruments of appropriate difficulty. As shown in Part III, data that was readily interpretable with the DLG was obtained using the FCI as pre-test and the CATS as post-test. Members of the PER/EER community are invited to share their experience with the DLG using the same or other test pairs.

HOW WERE THE INSTRUMENTS SELECTED? The choice of using FCI/CATS as a NIPPs-pair was made before the author joined the research group and is thus not part of this dissertation. However, the selection is justified first by the theoretical overlap in concepts on these two instruments, which is also noted by Pellegrino et al. (2013), who state that *"[...], both the Force Concept Inventory ([...]) and CATS ([...]) target multiple concepts associated with analyzing a system in static equilibrium (i.e., all forces balance, resulting in a motionless system)."* Second, it is justified by the fact that the CATS, unlike the FCI, is not a valid pre-test (and vice versa). The results of the revalidation confirm that the selection of CATS as the post-test (but not as the pre-test) was appropriate. The very linear relationship between FCI and CATS average scores makes them a suitable test pair to be analyzed with the DLG.

23.3 THE EFFECTIVENESS OF TUTORIALS

SUMMARY Using the DLG, the Tutorials (*Tut*) were found to be more effective in promoting conceptual understanding of statics than traditional instruction (*Trad*), i. e. instruction that did not make use of the materials. The *Tut* cohorts had a general post-instruction level of 12.3 ± 0.3 which is 2.9 ± 0.5 out of 27 points more than the *Trad* cohorts. The DLG lines do not intersect on the pre-test score range and the discriminative effect has a slightly larger estimated value for the *Tut* cohorts. It is hence concluded that students across all pre-test levels benefit from the Tutorials, and the effect may be even slightly stronger for the well-prepared (in terms of FCI score) students than

for those who initially have a very non-Newtonian belief system. The effects of other observed but uncontrollable variables such as class time, instructor, co-requisite physics courses, and the use of elements of JiTT could either be singled out and were then shown to be of substantially smaller magnitude compared to the effect attributed to the Tutorials, or worked towards a conservative estimation of the effectiveness of the Tutorials. The actual effect of the learning materials may thus be even stronger than the data suggests. Additionally, results from the reTest study show that the Tutorials are not only effective immediately after their implementation, but that the positive effect lasts throughout the students' further university studies (and possibly beyond).

The reTest study also revealed that the same level of understanding achieved by the *Tutorials* may be achieved by engagement as mechanics TA, likely because teaching requires many of the same core activities that are also promoted by the *Tutorials*, such as active engagement with the concepts in social interaction. Without the opportunity to practice these, student understanding is very likely to remain low until the end of their studies. Incorporating *Tutorials* into the instruction is an appropriate method to help students elevate their understanding *by the course* (rather than by a TA job) to a level higher than that achieved with traditional instruction only.

In future research, the effectiveness of the Tutorial worksheets in Kautz et al. (2018) on Mechanics of Materials, Kinetics, and Kinematics should also be evaluated, possibly in a similar manner (or adapted to the conditions where required).

CAN THE RESULTS BE INTERPRETED PRECISELY FOR ALL PRETEST SCORE LEVELS? The results of the revalidation study indicate that the CATS measurements are most precise for higher ability students in terms of the measured construct *conceptual understanding of statics*. As the CATS score is a proxy for ability, it can be expected that the measurements are most precise in the higher CATS score range (except the extreme scores). Judging by the overall positive slopes of the DLG lines, the students with higher abilities as measured by the CATS tend to have higher FCI scores. Therefore, (assuming constant precision of the FCI), the effect sizes at higher pre-test score levels (i. e. 70 %) are more reliable than those at lower pre-test score levels. This does not mean that the differences may not be interpreted for the other pre-test score levels. Despite the lower precision for the ability levels of the majority of *individual* students, the large sample sizes allow for highly precise *averages* to well interpret the shown differences by instructional method on all pre-test score levels.

ARE THE TUTORIALS TEACHING TO THE TEST? The Tutorials were partly developed based on the CATS results from the early tra-

ditional cohorts, which led to a few similarities between the situations inspected on the worksheets and the CATS items. Especially the worksheet on the most difficult concept (according to the data), *Static Equivalence*, is rather close to the respective CATS items. Looking through the lens of the situated cognition framework may lead to the conclusion that the apparent success of the Tutorials in fostering conceptual understanding could also be explained by the fact that the students in the *Tut* cohorts studied the concepts in a very similar context to the one on the measurement instrument. There is no evidence that they can transfer their understanding to other contexts. While it is true that there is no other quantitative evidence than the CATS data, evidence was presented that students are expected to answer such questions after instruction, as nicely illustrated by the following expert quote:

See also p. 104.

> *"[...] if they are not able to answer these questions at the end of the semester, then they should not even be able to take the exam, because it is so fundamental, [...]."*
>
> *"[...] wenn die das am Ende des Semesters diese Fragen nicht können, dann sollen sie bitte auch gar nicht zur Klausur können, weil das so basic ist, [...]."*
>
> *E8 #00:52:54-2#*

If these concepts are so basic, the students need to be given the opportunity to think about them thoroughly, especially because the concepts have also been shown to be difficult for students. The Tutorials provide this opportunity.

DOES THE IMPLEMENTATION OF THE TUTORIALS REQUIRE DEEP PEDAGOGICAL CONTENT KNOWLEDGE ON PART OF THE INSTRUCTORS? An essential aspect of successful learning materials is whether a regular instructor can implement the materials with the same effectiveness as the author of the materials or a person with a similar level of pedagogical content knowledge[2] required for the development of RBALM.

For this purpose, it is helpful to compare the use of research-based active-learning *materials*, e. g. the *Tutorials*, to active-learning *methods* such as *Peer Instruction* (Mazur, 1997), which makes use of response-systems (clickers) to pose in-class multiple-choice questions. Unlike the methods, for which the content (e. g. clicker questions) is generally designed by the individual instructor, the *Tutorials* (i. e. the published instructional materials) already include the relevant content in a way that has been informed by research. Finkelstein and Pollock (2005) describe the particular Tutorial tasks as "well documented and

2 as defined by Shulman (1986), pp. 9-10

easily implemented (in terms of both infrastructure and practice)". It seems therefore likely that materials such as the *Tutorials* are easier to implement than content-independent active-learning methods. The empirical data support this theoretical argument for Tutorial "transportability" or "exportability": Comparing the *TutOnly* cohorts taught by instructor B to the *TutJiTTl* cohorts taught by instructor A, who is one of the authors of the *Tutorials*, reveals no significant difference (see Figure 40), as it can be assumed that the use of JiTTl did not have a strong negative effect. Considering furthermore that the materials were implemented in the group recitation sessions, which are run by student TAs and not the course instructors (lecturers) themselves, an effect due to instructor would be even more surprising.

Instead, literature suggests that training of TAs in socratic dialogue is essential for the success of the *Tutorials* (e. g. Koenig et al., 2007). Instructors thus need to organize adequate training for the TAs that focuses on using socratic dialogue techniques and typical student thinking on each particular worksheet, rather than acquiring a level of expertise in (research on) pedagogical content knowledge similar to that of the material authors.

WHY ARE INSTRUCTORS OFTEN RELUCTANT TO IMPLEMENT THE TUTORIALS? The question may arise why the *Tutorials* are not used throughout the later cohorts. In the German higher education system, instructors are free to choose whichever method they like to teach "their" course. After instructor C filled the gap, instructor D took over and decided not to include the *Tutorials* in his teaching.

In spite of all the evidence for the effectiveness of Tutorials presented here and in previous studies, these materials are not as widely used as one would expect. In many cases, instructors are well informed about and willing to implement RBALM but run into situational barriers that must be overcome, (much like a lack of charging infrastructure can slow down or hinder the wide-spread use of electric mobility). Common reasons that are named by instructors for not using research-based methods and materials, even though they would like to, are lack of time or resources, class size, room layout, lack of departmental support, concern about incomplete coverage of the content, fear of bad student evaluations and possible consequences, and experience of student resistance against active learning (Henderson and Dancy, 2007).

Student resistance does indeed apply to Tutorials (e. g. reported by Finkelstein and Pollock (2005) or Riegler et al. (2016) and also experienced by the author herself), often because active engagement that challenges one's understanding is a strain. It may leave students with an unpleasant feeling of confusion and open questions, but it must not be forgotten that this is an indicator for learning, a process of conceptual change. Attending a lecture that follows a coherent flow

of thoughts is much easier for students to "digest". It often leaves students with the feeling that they learned more than in an interactive setting because the students may never (or only much later) become aware of their open questions (Deslauriers et al., 2019). Unfortunately, the depth of understanding is often an illusion (Rozenblit and Keil, 2002; Schwartz, 2013). Researchers suggest to remind students regularly about the mechanisms of learning and show evidence for the effectiveness of the chosen materials and methods (such as presented by this dissertation) to help increase student acceptance (e. g. Deslauriers et al., 2019).

In addition to student acceptance, the Tutorials require certain situational conditions such as a reasonable TA to student ratio, functional space, and adequately trained staff. In the investigated case, the TA to student ratio was not more resource-intensive than in the traditional settings (at worst 1 : 20, but often considerably better). The training of the TAs was minimal and incorporated in their weekly routine. The available space was adequate in that it allowed to rearrange tables and chairs to form group working stations.

Nevertheless, instructor D returned to a traditional setting after the Tutorials had been used - demonstrably with success - for several years. While the circumstances at the time indeed provided a reasonable excuse for the first year (short notice), and maybe the second year (restructuring other courses), it is clear by now that lack of time or resources is not the only obstacle.

It is a recognized phenomenon that STEM instructors (who are in many cases also active researchers in their own technical field) seem to measure by two standards regarding the value and necessity of evidence in technical fields and in education research (e. g. Hake, 2010, p. 24). While it is certainly not acceptable in their research community to make a claim about the effectiveness of any object of their research (e. g. a process, method or product) without presenting carefully evaluated evidence, this practice seems to be considered appropriate in case of teaching when instructors continue to claim that their students learn best the traditional way.

This attitude may be rooted in a common skepticism about the validity of education science and its theories and experiments as "real science" as illustrated by Heron (2018) who reports how her research group maneuvered through an area of conflict caused by the two different audiences of physicists and education/cognitive/learning scientists so that their work was taken seriously by their physics colleagues. They avoided statements about their theoretical framework in their publications and carefully selected terms and phrases that were acceptable in the physics community.

This double-standard attitude is also addressed by Freeman et al. (2014) who compare their results to standards from medical studies, concluding that if active learning were a medical treatment, tradi-

tional instruction would no longer be applied because of the existence of a demonstrably more effective treatment. While most instructors (with or without background in medicine) would agree that the employment of medical treatments requires evidence on effectiveness gathered in medical studies, many hold on to teaching strategies proven to be less effective, which seems irrational. Others are very enthusiastic about their teaching and often try new teaching strategies with the best intent for their students' learning, but fail to recognize the need for implementing evidence-based approaches. There are thus two possible underlying reasons for ignoring evidence: (1) The validity of the evidence is questioned because the validity of education science in general is questioned. (2) Seeing no necessity in presenting evidence for the effectiveness of instructional strategies has the consequence that, when evidence is presented, it is not valued.

The phenomena described above contribute to a situation where certain ideas about Tutorials persevere, e. g.

- *"The Tutorials help only the weak students"*. The results presented here contradict this statement. The Tutorials were shown to increase the average score for all levels of prior understanding, but especially the strong students. The idea may result from the rather small steps taken in the worksheets or from the belief that the performance of a group will be the average performance of its individual members such that "strong" students are slowed down while "weak" students benefit. This belief may be correct in some cases, but often, the group enables processes that individuals people cannot perform on their own. Group discussions, for example, can trigger new thoughts, questions asked by a (supposedly) weak member of a group can challenge understanding of which also the (supposedly) strong members of a group benefit.

- *"The Tutorials do not help students solve quantitative problems" / "The time should better be spent on more problem-solving"*. Quantitative problem-solving is essential for engineers, but an incorrect quantitative result is worthless, and results are highly likely to be incorrect if they are based on incorrect assumptions or reasonings. "More of the same" activities only help to a certain degree to avoid such errors. Understanding the rationale behind the systematic problem-solving algorithms and procedures on a conceptual level helps students move from the novice to an expert status. Lacking a grasp of the underlying concepts of the systematic problem-solving algorithms and procedures, however, can be dangerous as errors remain undetected.

- *"In the Tutorial setting, the instructor cannot control if students leave the group discussions with wrong ideas in their heads."* No, they cannot. But this is equally impossible in a teacher-centered setting,

likely even more so because of unidirectional communication, which is broken up in the Tutorial setting. Here, the instructor or TA listens in on the students' discussions from time to time and many worksheets even include checkpoints. However, the paradigm of the Tutorials acknowledges that students leave the classroom and do not have all addressed concepts straight in their heads, and that it takes time to build understanding. Its aspiration is to start the process of struggling for understanding in a constructivist manner.

- *"There should be worked-out solutions to the worksheets that can be given to the students."* The motivation for this statement is often rooted in the concern discussed in the previous point. However, general disadvantages of worked-out solutions are that they create the illusion of understanding. Furthermore, their availability is a temptation to give up thinking and discussing earlier than one would do without. Most importantly, however, worked-out solutions are inadequate for Tutorials because they shift the focus away from the process towards the product.

These ideas about Tutorials in particular hint at the existence of strongly held beliefs (and maybe even misconceptions) about teaching and the mechanisms of learning in general that are in conflict with the constructivist viewpoint. It seems that these fundamental conceptions must be changed before the Tutorials can be accepted. For this purpose, future research should investigate the question how to best disseminate RBALM in general and the Tutorials in particular.

FINAL CONCLUSION Under the assumption that conceptual understanding is considered an important learning objective in engineering mechanics, the results presented in this dissertation confirm that *Tutorials*, as a form of research-based active learning materials, do in fact result in improved quantifiable short- and long-term understanding under realistic conditions with minimal additional effort and resources. It was also shown that instructors other than the *Tutorial* authors can successfully implement these instructional materials. Training of the teaching assistants in Socratic dialogue may even further increase the effectiveness.

APPENDIX

A

THE CATS

The following version of the CATS was created as supplemental material to this dissertation. It serves as a reference for readers who are encouraged to consult it while reading the dissertation, as the items are not reprinted every time they are addressed.

> PLEASE NOTE: While the developers have decided to share the CATS for research purposes, care must be taken not to invalidate the instrument. As tempting as it may be, the items must *never* be used as learning material or in exams (even if the CATS is not administered for research in the same course). Such practice would sooner or later lead to a teaching-to-the-test situation and hence invalidate the instrument.

Note that there are two different orders of the CATS: the standard order, which is administered to the participants, and the concept order, which groups the items by concept and is often used for analysis of the test. The following pages show the items in standard order as the students would see them, but with the following modifications:

- Each page shows one item in *both* languages (but neither the figures nor the font have been altered for this document, they do have a slightly different appearance in each version).
- On the top of each page, the item numbers in both orders, standard and concept, were added.

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 1 | Concept: 10

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 2 | Concept: 1

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 3 | Concept: 13

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 4 | Concept: 4

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 5 | Concept: 19

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 6 | Concept: 22

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 7 | Concept: 7

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 8 | Concept: 16

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 9 | Concept: 2

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 10 | Concept: 11

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 11 | Concept: 25

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 12 | Concept: 14

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 13 | Concept: 23

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 14 | Concept: 21

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 15 | Concept: 3

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 16 | Concept: 5

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 17 | Concept: 17

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 18 | Concept: 8

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 19 | Concept: 12

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 20 | Concept: 26

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 21 | Concept: 15

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 22 | Concept: 24

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 23 | Concept: 20

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 24 | Concept: 6

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 25 | Concept: 18

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 26 | Concept: 27

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

Standard: 27 | Concept: 9

The CATS is available in a separate password-protected file at https://doi.org/10.15480/336.3504.

Please contact the author (jdirenga@posteo.de) or the editor (kautz@tuhh.de) for access.

B

ELEMENTS IN CATS ITEMS

Element	Items by concept																									
	2	9	15	4	16	24	7	18	27	1	10	19	3	12	21	8	17	25	5	14	23	6	13	22	11	20	26
rollers						x				x	x	x		x										x			
supports				x	x	x				x	x	x	x	x	x	x	x		x	x	x						
support reactions																		x	x						x		
connecting pins				x	x	x					x																
ropes	x		x												x				x	x							
bodies	x	x	x				x	x	x	x					x				x	x	x	x	x	x			
pin in slot					x								x	x	x												
springs												x	x			x	x										
forces (directed)								x	x	x		x	x	x	x	x	x					x	x	x	x	x	x
forces (arbitrary)				x	x	x	x											x	x	x	x						
forces (contact)		x															x										
forces (weight)	x	x	x												x							x	x				
forces (rope)	x		x																	x							
couples							x	x	x				x	x	x	x	x	x	x	x	x				x	x	x
no friction																x	x	x									
friction coefficient										x	x	x										x	x	x			
free-body diagram	x	x	x																x	x	x						
"freigeschnitten"	x		x																								
angles										x	x	x	x	x	x	x	x			x	x				x		
welded joint																					x						
hydraulic cylinder																											
points, distances							x	x	x		x																

Elements found in each CATS item, based on Geier (2016)

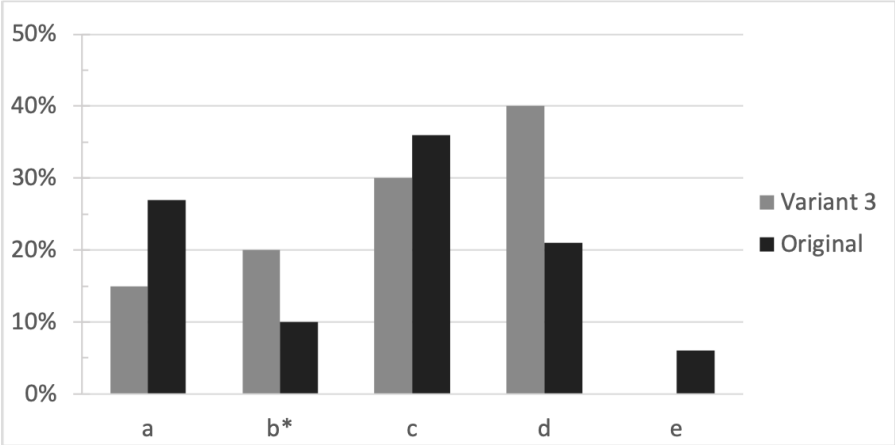
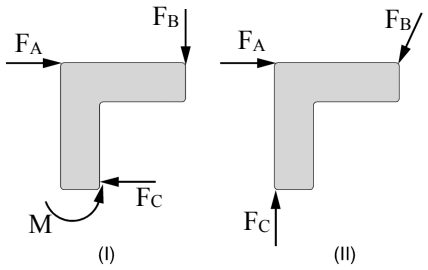


Figure 48: Response distribution on variant 3 of item 20, compared to the original version.

Version A

The forces and couples on the L-shaped body act at the points indicated. Directions and senses are represented accurately by the arrows, and all drawn forces and couples are unequal to zero.



Assuming the magnitudes of the forces and couple have the right values, could these bodies be in equilibrium? You may introduce quantities as needed. Show your work.

a.) How do the forces and moments acting on body (I) have to be chosen to achieve equilibrium?

→ Can body (I) be in equilibrium under the given conditions?
Yes ☐ No ☐

b.) How do the forces and moments acting on body (I) have to be chosen to achieve equilibrium?

→ Can body (II) be in equilibrium under the given conditions?
Yes ☐ No ☐

Optional:

Midterm grade: _____

☐ did not take part in midterm because _____

☐ I participated in an interview and I already saw this (or a similar) question there.

30.01.19

Figure 49: English translation of variant 3 of item 20

D

FACTOR ANALYSIS

Table 16: Factor loadings of noBlanks data set. (*)Item omitted.

Item	Drawing forces	Newton's Third Law	Static equiv.	Roller joint	Pin-in-slot joint	Frictionl. contact	Repres. loads*	Limits on friction force
2	0.71							
9	0.44							
15	0.69							
4		0.58						
16		0.59						
24		0.56						
7			0.39					
18			0.37					
27			0.36					
1				0.61				
10				0.44				
19				0.59				
3					0.43			
12					0.47			
21					0.56			
8						0.42		
17						0.52		
25							0.44	
5								
23							0.44	
14								
6								0.49
13								0.71
22								0.38
11								
20*								
26								

Table 17: Factor loadings of Trad data set. (*)Item omitted.

Item	Drawing forces	Newton's Third Law	Static equiv.*	Roller joint	Pin-in-slot joint	Frictionl. contact	Repres. loads*	Limits on friction force
2	0.68							
9	0.40							
15	0.71							
4		0.62						
16		0.48						
24		0.37						
7			0.48					
18								
27*								
1				0.59				
10				0.35				
19				0.53				
3					0.50			
12					0.43			
21					0.34			
8						0.31		
17						0.57		
25							0.46	
5								
23							0.60	
14								
6								0.42
13								0.66
22								0.38
11								
20*								
26								

Table 18: Factor loadings of Tut data set. (*)Item omitted.

Item	Drawing forces	Newton's Third Law	Static equiv.	Roller joint	Pin-in-slot joint	Frictionl. contact	Repres. loads*	Limits on friction force
2	0.66							
9	0.48							
15	0.60							
4		0.58						
16		0.59						
24		0.51						
7			0.53					
18			0.47					
27							0.35	
1				0.50				
10				0.42				
19				0.60				
3					0.43			
12					0.55			
21					0.39			
8						0.64		
17						0.32		
25							0.52	
5								
23							0.68	
14								
6								0.61
13								0.68
22							0.39	
11								
20*								
26							0.46	

E

CATEGORICAL JUDGEMENT SCHEME

Table 19: Categorical judgement scheme (Jorion et al., 2015, Table 11). Numbers in brackets indicate numbers of items which may fall outside of this recommendation.

Analysis	Excellent	Good	Average	Poor	Unacceptable
CTT					
Item stats.					
-Difficulty	.2 to .8	.2 to .8 (3)	.1 to .9	.1 to .9 (3)	.0 to 1.0
-Discrimination	>.2	>.1	A >.0	>.-2	>.-1.0
Total score reliability					
-Cronbach's α total	>.9	>.8	>.65	>.5	>.0
-Cronbach's α_{-i}	$\forall\alpha_{-i} : \alpha_{-i} < \alpha$	(3)	(6)	(9)	>(9)
IRT					
Item measures					
-All items fit model	(2)	(4)	(6)	(8)	(10)
Structural analysis					
Exploratory FA	Conforms to predicted constructs	(5)	(10)	(15)	>(15)

F

IRC DISTRACTOR ANALYSIS

F.1 DRAWING FORCES CONCEPT

Among the items of the FBD concept, item 9 seems to be the most difficult but also the least discriminating item, while item 15 is the easiest and most discriminating one. The number of blank responses remains moderate as none of the items comes very late.

Item 2: Item 2 is rather easy with a reasonable discrimination. There are nearly no blank responses. The distractors all follow a similar pattern, in that their attractiveness decreases with increasing test score. The ranking of the attractiveness of the distractors for low test scores is quite stable until at about 10 points, the ranking changes. Distractor (e) (drawing the wrong type of force or omitting weight of external body) is generally ineffective. Distractor (b) (including forces acting on external bodies) remains also below the pure guessing probability of 0.2. While distractors (c) (drawing internal forces) and, to a lesser extent, (a) (drawing internal forces and including forces acting on external bodies) are very attractive for low scoring students, the attractiveness decreases fast with increasing score, and distractor (b) becomes equally or even more attractive, but here, the amount of incorrect responses is already very low.

Item 9: The low scoring students are about equally attracted to distractor (a) (including forces acting on external bodies) and (d) (drawing upward internal forces), though the latter declines more rapidly with increasing test score than the former. Distractor (b) (drawing the wrong type of force or omitting weight of external body) is largely and distractor (c) (drawing downward internal forces) is completely ineffective.

Item 15: The distractors are only relevant for the lower third of the score range. Distractor (b) (drawing upward internal forces) is most attractive for the low scoring students, but all distractors (including blank responses) are close to the guessing probability of 0.2.

F.2 NEWTON'S THIRD LAW CONCEPT

Item 4: There is no clear dominance but a slight preference for distractor (e) 2Force.

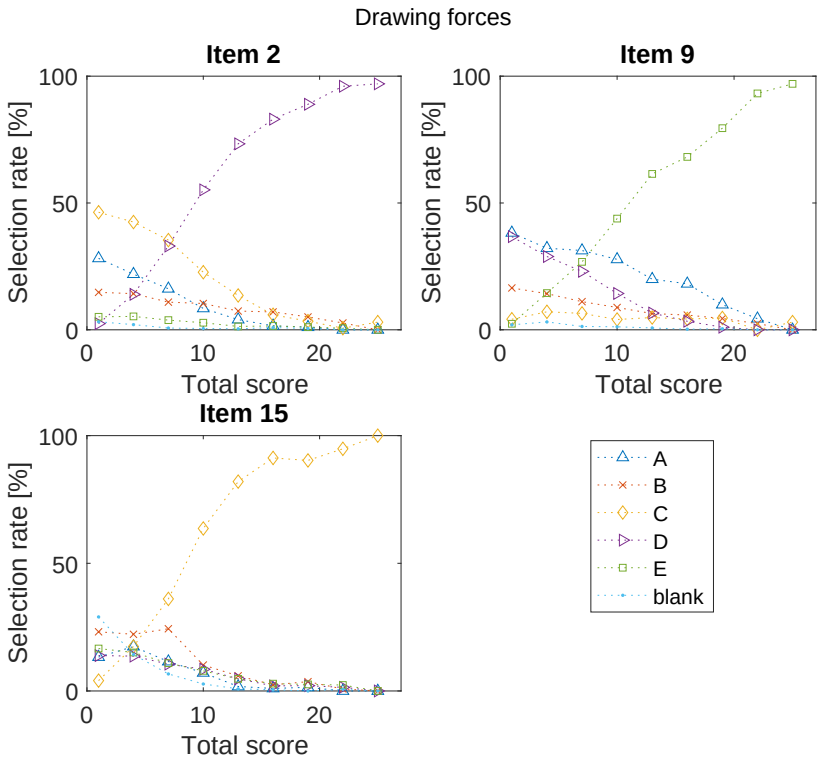


Figure 50: IRC of items on the Drawing forces concept category

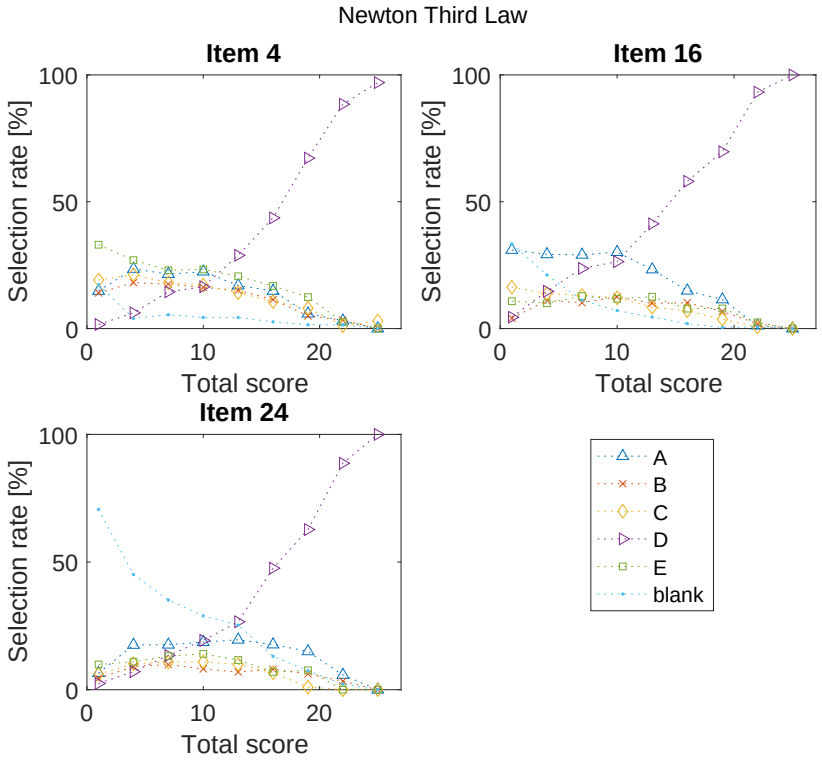


Figure 51: IRC of items on the Newton's Third Law concept category

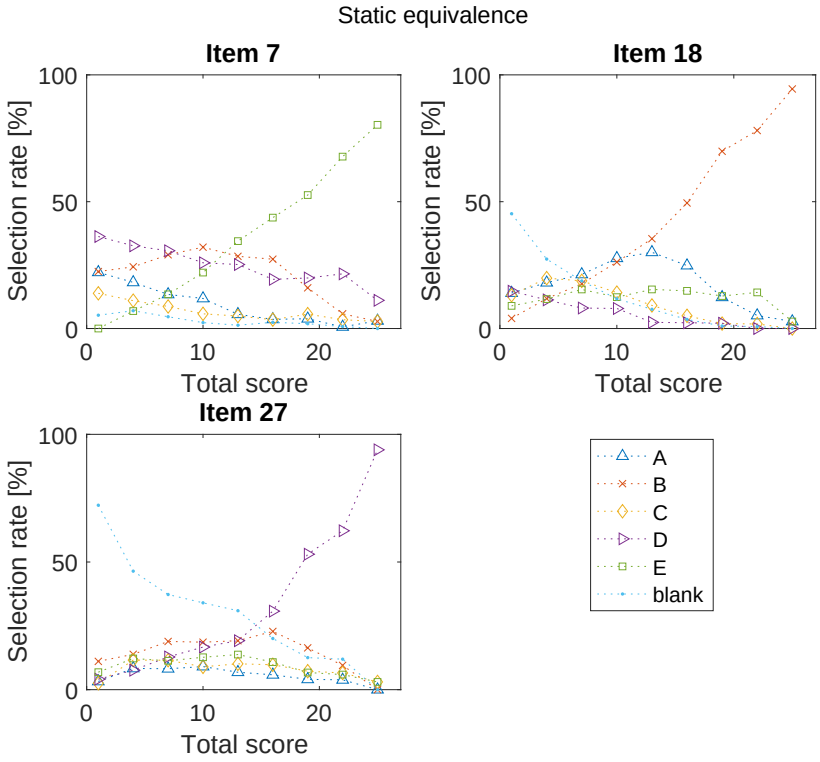


Figure 52: IRC of items on the Static equivalence concept category

Item 16: In case of item 16, the 2Force distractor is clearly dominant and also quite stable over the lower part of the score scale. The other distractors are constantly below the guessing probability.

Item 24: Blank responses are dominating the late item 24. Low scoring students probably are too slow to even view the item. This results in a very characteristic shape of the distractor curves with the maximum distractor attractiveness not located at the low end but rather in the middle of the score range. In accordance with the other items, distractor (a) (2Force) is slightly more attractive that the others.

F.3 STATIC EQUIVALENCE CONCEPT

Item 7: Two distractors are dominant, but in different regions of the score scale: While distractor (d) (CentrM) is most attractive to both, the very low and very high scoring students, distractor (b) (M=F) peaks in the middle range of the scores.

Item 18: In accordance with item 7, the most attractive distractor on item 18 is distractor (a) (M=F), at least for the middle range of scores. In the very low range, the blank response option is dominant, and in the very high score range, distractor (e),

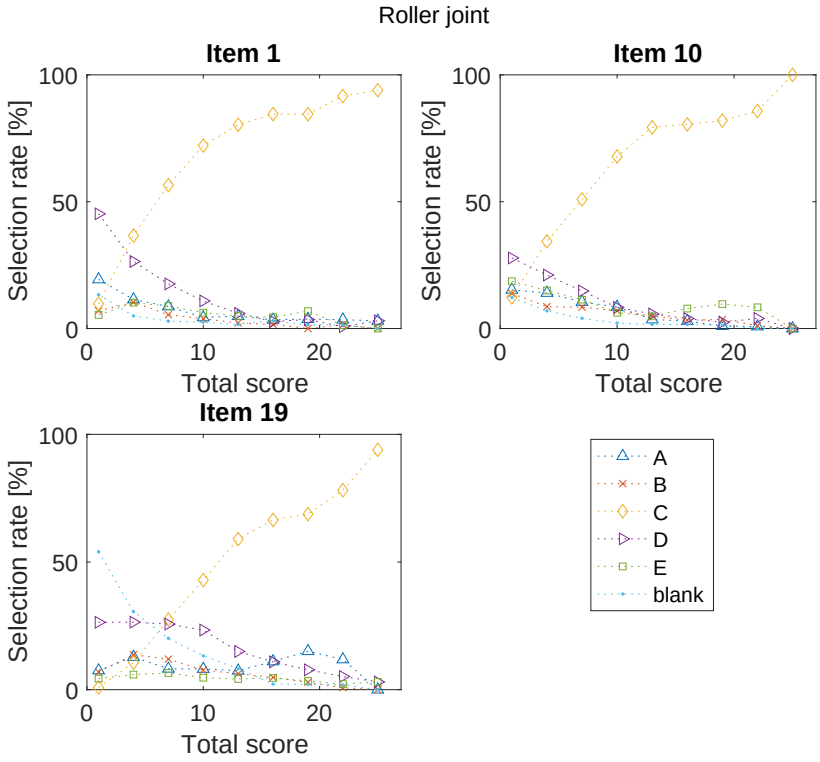


Figure 53: IRC of items on the Roller joint concept category

which is very similar to the correct response, seems to be more than or at least equally attractive as distractor (a).

Item 27: Item 27 is the last item on the test and is therefore affected strongest by the time limit. In light of the vast amount of blank responses, the distractors play a subordinate role. The characteristic "arching" curves caused by the blank responses are clearly visible. Nevertheless, distractor (b) (M=F) is most attractive, again in accordance with items 7 and 18.

F.4 ROLLER JOINT CONCEPT

Item 1: The preferred distractor is (d) (2Force, force acts along the arm supporting the roller).

Item 10: Item 10 exhibits similar IRCs, but with a less dominant distractor (d) (direction of internal forces, similar to ApplF) in the low score range. In the high score range, the attractiveness of distractor (e) ($\arctan(\mu)$) is slightly elevated.

Item 19: In the lower part of the score range, item 19 is dominated by distractor (d) (2Force), in accordance with item 1. In the higher score range, distractor (a) ($\arctan(\mu)$) is more attractive, in accordance with item 10.

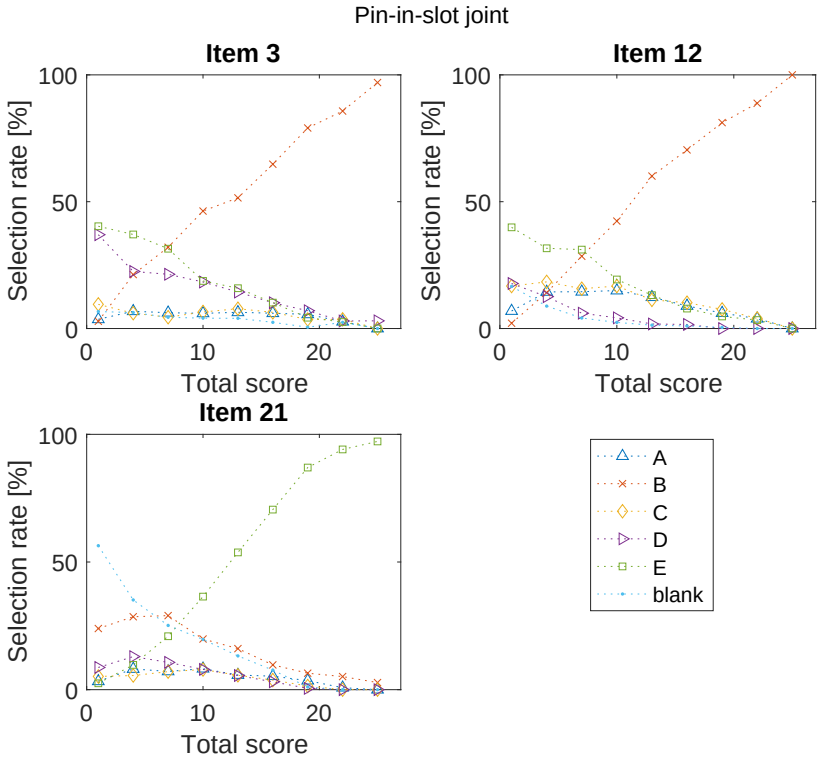


Figure 54: IRC of items on the Pin-in-slot joint concept category

F.5 PIN-IN-SLOT JOINT CONCEPT

Item 3: Distractor (e) (Moment) dominates in the low test score range, followed by distractor (d) (Motion). The other distractors are generally ineffective.

Item 12: Item 12 is dominated by distractor (e) (Moment) in the lower score range.

Item 21: Item 21 is influenced by blank responses resulting in the characteristic maximum of distractor attractiveness at around 5 points on the score scale. Distractor (b) (Moment) is the only attractive distractor that stands out over the entire score range.

F.6 FRICTIONLESS CONTACT CONCEPT

Item 8: Item 8 seems to be rather difficult. The dominant distractor is (c) (only case (II) is possible), which remains quite attractive far into the upper half of the score range, while distractor (b) (only case (I) is possible) is only attractive for low scoring students. The preference for distractor (c) can be explained by the torque driving the arm such that the force perpendicular to the arm indicates the direction of movement (if the arm were free to move).

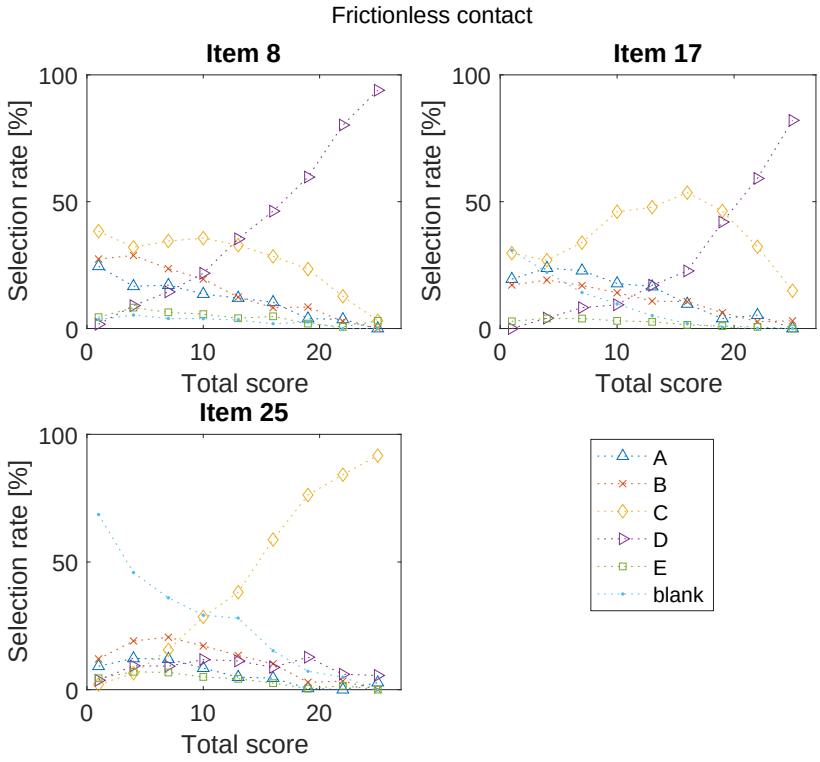


Figure 55: IRC of items on the Frictionless contact concept category

Item 17: This item is very difficult as can be seen from the late rise of the correct response curve (d) (moment and non-normal force both impossible). Apart from the "not enough information" distractor (e), the distractors are about equally attractive for the very low test scores. With increasing test score, most distractors become less attractive but distractor (c) (non-normal force is possible) becomes more attractive in the middle of the test score range. This pattern indicates that low scoring students tend to guess while higher scoring students understood at least that moments are impossible to transmit in such a setup. Only the very high scoring students see that the force has a tangential component and therefore is impossible if friction is neglected.

Item 25: Item 25 is dominated by blank responses. A slight preference for distractor (b) (moment possible, force impossible) in the low score range and for distractor (d) (both impossible) in the high score range can be detected. The higher scoring students who miss this item seem to have at least understood that the moment is impossible. Identifying the force as possible seems to be the most difficult part on this task. As in case of the Slot concept, it is possible that forces/couples were confused with possible direction of movement.

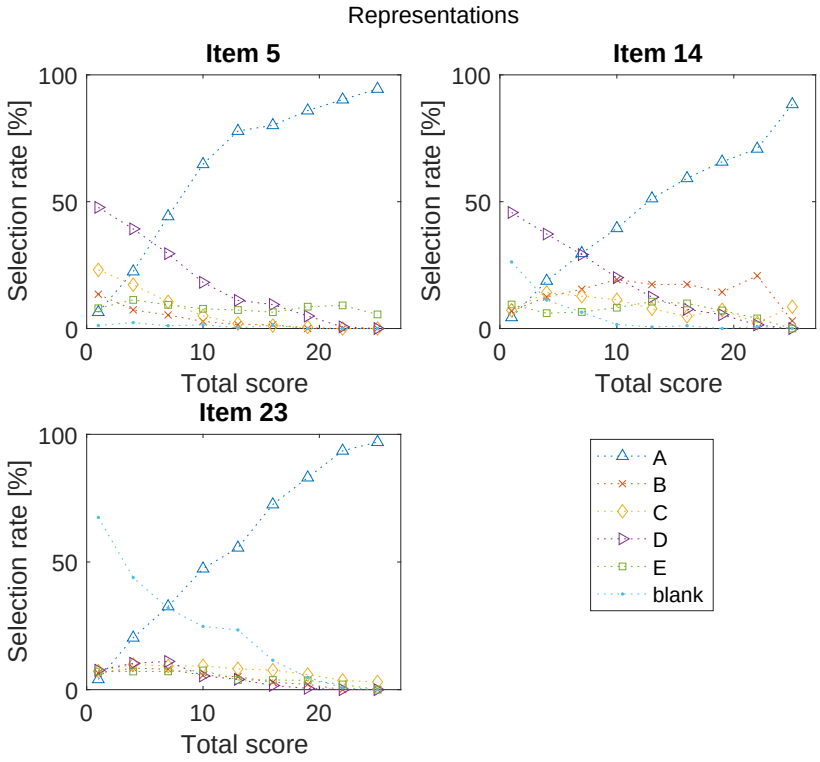


Figure 56: IRC of items on the Representations concept category

F.7 REPRESENTATIONS CONCEPT

Item 5: Item 5 seems to be rather easy but with good discrimination. Distractor (d) (two forces, one couple) is the dominant distractor in the low and middle score range. Distractor (e) (one force at 45 degrees) is independent of test score in its attractiveness such that it becomes the dominant distractor in the very high score region. Many of the interviewed experts misinterpreted the point of interest to be the force of the rope on the ground support. This misunderstanding might result from the description in the German translation "lower-left corner of the body", which might be carelessly misinterpreted as the left-most corner of the entire setup. For this scenario, distractor (e) would actually be the correct response.

Item 14: Item 14 discriminates poorly, as judged by the slope of the correct response curve. In the low score range, distractor (d) (force and couple) is very attractive, while distractor (b) (force at arbitrary angle) is more constantly attractive and dominates the high score range. The conception that the horizontal and vertical component of the force are independent is rather insignificant (distractors (c) and (e)).

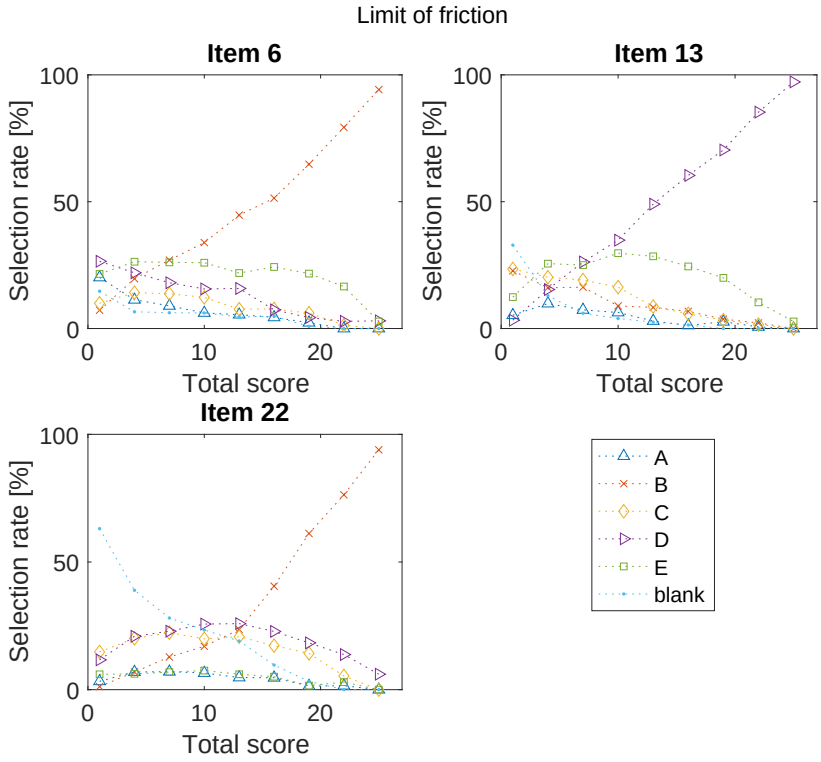


Figure 57: IRC of items on the Limit of friction concept category

Item 23: Item 23 does not discriminate well, either. Apart from the blank response option, the distractors are all equally unattractive, with a slight preference for (c) (no moment) in the higher score range. This might be influenced by the joint connecting the system to the ground, which cannot exert a moment.

F.8 LIMIT OF FRICTION CONCEPT

Item 6: Item 6 does not discriminate well. Students in the higher score range clearly prefer distractor (e) (MuN). In the lower score range, the dominance is not as clear because distractor (d) is also chosen frequently. This distractor results from the difference between the force applied to the middle block and the one applied to the upper block, which resembles applying the equilibrium condition for horizontal forces but including a force acting on an external body.

Item 13: A similar pattern shows in item 13, also with distractor (e) (MuN) being very attractive, reaching far into the higher score range. In the lower score range, the difference between the attractiveness of distractors is not so strong, with (c) (F-MuN) being the second most frequently chosen distractor. Distractor (b) (MuN with weight force of upper block omit-

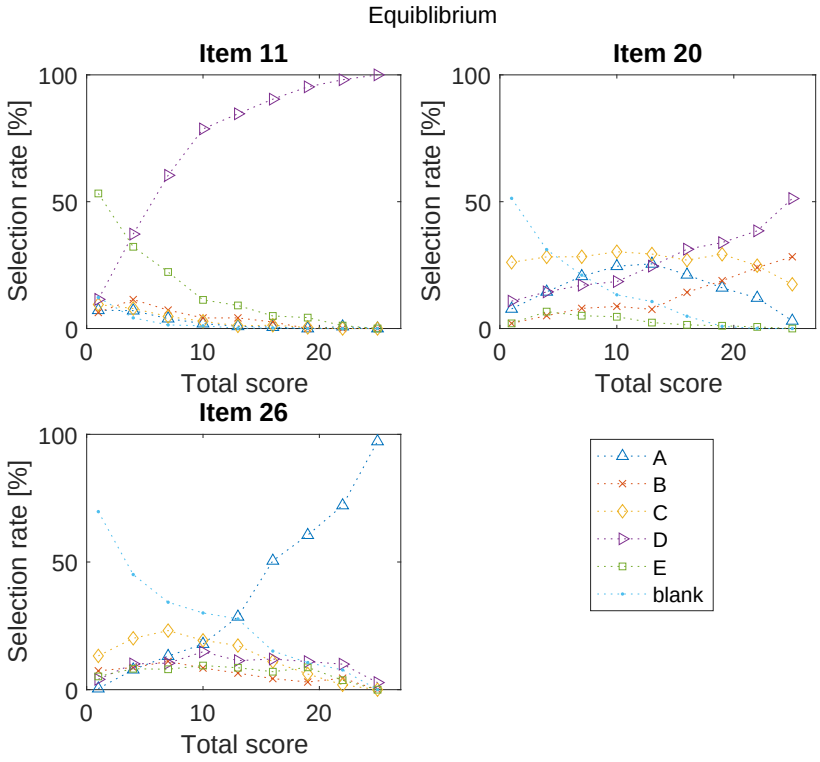


Figure 58: IRC of items on the Equilibrium concept category

ted) is similarly attractive. Distractor (a) (F-MuN with weight force of upper block omitted) is rather ineffective, probably because it is a combination of three mistakes.

Item 22: Item 22 discriminates well between the medium and the high-scoring students. Blank responses dominate the lower score range. Distractor (d) (limit of friction) is most effective, followed by (c) which resembles misconception F-MuN. Students may also select this response if MuN is combined with a calculation error (20/4 instead of $20 \cdot 0.4$). Distractors (a) (F-MuN, ignoring symmetry) and (e) (MuN, ignoring symmetry) are ineffective which is evidence that the consequences of the symmetric setup are well interpreted.

F.9 EQUILIBRIUM CONCEPT

Item 11: Item 11 is a very easy item. The only attractive distractor is (e) (force at 20 degrees plus a moment).

Item 20: Item 20 is clearly very difficult. Even students in the higher score range only answer this item correctly in less than one out of three cases. Judging only by the patterns of the ICCs, the correct response seems to be the distractor (d) (only body II is in equilibrium), because it is monotonously increasing

and the most attractive option in the high score range, but this is incorrect. The correct response is (b) (both not in equilibrium) whose ICC has a similar shape as the ICC of distractor (d), but is overall less attractive. In the lower half of the score range, the correct response is even as unattractive as the generally ineffective distractor (e) (not enough information), which is far less than guessing probability.

The pattern of the distractors is also very diverse: Distractor (c) (body I in equilibrium, body II not) is more or less constantly attractive throughout the score range, distractor (a) (both in equilibrium) has its peak of attractiveness in the medium score range, and distractor (d) increases in attractiveness with increasing total score. This leads to the following pattern: in the medium range, distractors (a), (b), and (c) are almost equally attractive, in the low range, (c) dominates, and in the high range, (d) dominates. The low scoring students tend to respond blank or miss the equilibrium of forces (body I in equilibrium, body II not), while the high scoring students chose this response equally often, but even more frequently miss the equilibrium of moments (body I impossible, body II possible).

Item 26: Item 26 is dominated by blank responses. Distractor (c) (counter-clockwise couple in R: imbalance of vertical forces remains, and imbalance of moments is introduced) is preferred in the lower score range and distractor (d) (upward force in Q: imbalance of vertical forces is equalized, but imbalance of moments is introduced) in the upper range. Distractors (b) (horizontal force in P: imbalance of vertical forces remains) and (e) (clockwise couple in Q: imbalance of vertical forces remains, and imbalance of moments is introduced) are equally unattractive.

Table 20: Response distribution in percent for the standard sample. The correct answer is marked with an asterisk.

	Responses [%]					
Item	A	B	C	D	E	blank
1	7	4	*67	12	7	3
2	10	10	24	*53	3	1
3	6	*44	6	18	22	4
4	19	15	16	*25	22	4
5	*60	4	7	21	8	1
6	7	*38	11	16	24	5
7	11	28	7	24	*27	3
8	13	18	32	*26	6	4
9	24	9	5	15	*46	1
10	8	7	*63	11	9	3
11	3	4	3	*73	16	1
12	13	*44	14	5	20	3
13	6	11	14	*39	25	5
14	*41	17	10	21	8	4
15	7	13	*60	7	8	4
16	24	10	11	*35	11	9
17	18	14	42	*13	3	10
18	23	*31	13	6	14	13
19	9	8	*42	21	5	15
20	21	*9	28	23	4	15
21	6	20	6	8	*41	20
22	6	*22	20	23	6	23
23	*47	5	8	6	6	27
24	17	8	9	*26	11	29
25	8	16	*31	9	5	31
26	*25	7	18	11	8	31
27	8	18	9	*21	11	34

Table 21: Response distribution in percent of a US sample (Román et al., 2010a, converted from concept order to standard order). The correct answer is marked with an asterisk.

	Responses [%]				
Item	A	B	C	D	E
1	9	5	*68	11	7
2	12	11	23	*51	3
3	5	*70	4	15	6
4	19	26	13	*25	16
5	*63	4	5	22	5
6	4	*29	16	14	37
7	10	30	8	21	*32
8	13	16	40	*28	2
9	23	12	8	15	*42
10	4	5	*78	5	8
11	7	8	8	*62	15
12	6	*70	12	5	7
13	2	9	13	*29	47
14	*52	18	6	16	8
15	8	15	*59	11	7
16	29	8	11	*44	8
17	16	11	45	*26	2
18	30	*35	11	9	15
19	9	5	*58	23	5
20	28	*16	26	29	1
21	4	14	2	7	*74
22	7	*33	14	41	5
23	*71	4	9	9	6
24	25	11	13	*33	18
25	18	14	*58	7	3
26	*49	3	16	22	11
27	11	22	13	*37	17

G

SUGGESTIONS FOR IMPROVEMENT OF THE CATS

The interviews with experts and students revealed a few minor issues with some CATS items. The following is a collection of suggestions for improvement of the CATS. They are listed in no specific order.

TRANSLATION Generally, the translation should be revisited to remove any unnecessarily complicated phrasing. Specifically, the phrasing in the Pin-in-slot items should be aligned among each other. In addition, it is suggested that the translation of item 20 is adapted to be closer to the original to make sure that non-positive magnitudes for forces are more easily identified as a violation of the equilibrium conditions.

ITEM 25 The term used in the response options of item 25 should match the term used in the question, i. e. the term "Reaktionen" (engl. "reactions") should be substituted by the more specific term "Lagerreaktionen" (engl. "support reactions") to avoid any confusion with reactions in terms of motion.

ITEM 8 Both, the text and the illustrations asking whether the forces are possible were difficult to interpret and should be redesigned. From the text, one student initially interpreted (I) to show the lever ("Kipphebel") and (II) to show the arm ("Schwenkarm"). To avoid this misunderstanding, the reference to the images (I)/(II) should be placed *after* the words they refer to, i. e. "Kann die Kraft auf den Kipphebel *parallel* (I) bzw. *senkrecht* (II)..." In the illustrations, the lever as the direction-giving object (parallel or perpendicular) is not shown together with the force. Instead of showing only one part of the setup (the rocker) and the force acting on it, the interacting bodies could both be shown separately with the force pair acting between them.

ROLLER ITEMS 1 AND 10 As the bodies are not marked with weights, according to the instructions, their weight force should be neglected, resulting in an unstable system. Adding a force that acts downward on the block and the platform, respectively, would solve this issue.

TEXT/IMAGE MISMATCH In item 1, the term "große Rolle" (large roller) is a legacy from a prior version in which the size of the rollers on the side was much smaller. It should be changed to "untere Rolle"

(lower/bottom roller) as the difference in size is not that significant anymore.

In item 4, the term "nebenstehenden" should be changed to "folgenden", as it refers to an incorrect position of the images showing the distractors.

HATCHES In items 1, 2, 4, 12, and 25, hatches illustrating fixed structures are missing or a very light and thus barely visible print. These should be added or enhanced for clarity.

FORCES AS VECTORS, NOT COMPONENTS It may be necessary to add a hint to the instructions that the items work with forces as vectors, not force components.

BIBLIOGRAPHY

ABET. History. http://www.abet.org/about-abet/history/, last accessed in March 2020. (Cited on page 10.)

Adams, W. K. and Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9):1289–1312. (Cited on pages 41 and 79.)

Ambrose, B. S. (2014). PER-based Tutorials for quantum physics. https://www.compadre.org/per/items/detail.cfm?ID=13266. (Cited on page 26.)

Ambrose, B. S. and Wittmann, M. (2007). Intermediate mechanics Tutorials. http://faculty.gvsu.edu/ambroseb/research/IMT.html. (Cited on page 26.)

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C. (Cited on pages 39, 40, 67, and 69.)

Anderson, E., Taraban, R., and Hudson, D. (2009). A study of the impact of visuospatial ability, conceptual understanding, and prior knowledge upon student performance in engineering statics courses. pages 14.119.1–14.119.10. (Cited on pages 63 and 178.)

Andrews, T. M., Leonard, M. J., Colgrove, C. A., and Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE-Life Sciences Education*, 10(4):394–405. (Cited on pages 2, 4, 24, and 187.)

Arons, A. B. (1981). Thinking, reasoning and understanding in introductory physics courses. *The Physics Teacher*, 19(3):166–172. (Cited on page 25.)

Arons, A. B. (1990). *A guide to introductory physics teaching*. John Wiley & Sons, New York. (Cited on page 25.)

Arora, M. L., Rho, Y. J., and Masson, C. (2013). Longitudinal study of online statics homework as a method to improve learning. *Journal of STEM Education: Innovations and Research*, 14(1):36–44. (Cited on page 65.)

Atadero, R. A., Rambo-Hernandez, K. E., and Balgopal, M. M. (2015). Using social cognitive career theory to assess student outcomes of

group design projects in statics. *Journal of Engineering Education*, 104(1):55–73. (Cited on page 219.)

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD, 2nd edition. (Cited on pages 90 and 161.)

Beichner, R. (2009). An introduction to physics education research. In *Getting Started in PER*, volume 2. (Cited on pages 4 and 187.)

Benegas, J. and Flores, J. S. (2014). Effectiveness of Tutorials for introductory physics in Argentinean high schools. *Physical Review Special Topics - Physics Education Research*, 10(1):010110. (Cited on pages 35 and 70.)

Bernhard, J. (2000). Teaching engineering mechanics courses using active engagement methods. In *Physics Teaching in Engineering Education (PTEE 2000), 13-17 June 2000, Budapest, Hungary*. (Cited on page 20.)

Blanton, P. (2001). Lessons learned from Arnold Arons. *The Physics Teacher*, 39(8):506–507. (Cited on page 25.)

Boles, W., Goncher, A., and Jayalath, D. (2015). Uncovering misconceptions through text analysis. In *6th Research in Engineering Education Symposium*, Dublin, Ireland. (Cited on page 40.)

Bond, L. (2005). Carnegie Perspectives: Who has the lowest prices? http://archive.carnegiefoundation.org/perspectives/who-has-lowest-prices. (Cited on pages 184 and 190.)

Bortz, J. and Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Springer, Berlin Heidelberg, 7th edition. (Cited on page 94.)

Bransford, J. D., Brown, A. L., and Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* National Academy Press. (Cited on pages 1, 2, 178, and 190.)

Brommundt, E., Sachs, G., and Sachau, D. (2007). *Technische Mechanik: eine Einführung*. Oldenbourg, München, 4th edition. (Cited on pages 3, 12, 14, 118, and 119.)

Brose, A. and Kautz, C. (2011). Identifying and addressing student difficulties in engineering statics. In *Proceedings of the 2011 ASEE Annual Conference and Exposition*, Vancouver. (Cited on pages 4, 20, 30, and 116.)

Brown, S., Lutz, B., Perova-Mello, N., and Ha, O. (2019). Exploring differences in statics concept inventory scores among students and practitioners. *Journal of Engineering Education*, 108(1):119–135. (Cited on pages 65 and 66.)

Brown, S., Montfort, D., Perova-Mello, N., Lutz, B., Berger, A., and Streveler, R. (2018). Framework theory of conceptual change to interpret undergraduate engineering students' explanations about mechanics of materials concepts: conceptual change in mechanics of materials. *Journal of Engineering Education*, 107(1):113–139. (Cited on page 20.)

Bundesministerium für Bildung und Forschung (BMBF) and Kultusministerkonferenz (KMK) (2013). DQR - German EQF Referencing Report. Technical report. (Cited on page 216.)

Burkhardt, P. J. (2015). The Effect of Additional Statics Class Time on At-Risk Student Performance. In *122nd ASEE Annual Conference & Exposition*, Seattle, WA. (Cited on page 226.)

Chi, M. T. H. (2013). Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes. In *International handbook of research on conceptual change*. Routledge. (Cited on page 2.)

Cho, E. (2016). Making reliability reliable: a systematic approach to reliability coefficients. *Organizational Research Methods*, 19(4):651–682. (Cited on page 86.)

Close, H. G., Gomez, L. S., and Heron, P. R. L. (2013). Student understanding of the application of Newton's second law to rotating rigid bodies. *American Journal of Physics*, 81(6):458. (Cited on page 33.)

Cochran, M. J. and Heron, P. R. L. (2006). Development and assessment of research-based tutorials on heat engines and the second law of thermodynamics. *American Journal of Physics*, 74(8):734–741. (Cited on page 33.)

Coe, R. (2002). It's the effect size, stupid; What effect size is and why it is important. In *British Educational Research Association annual conference*, Exeter. (Cited on page 198.)

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Assoc Inc, Hillsdale, N.J, 2nd edition. (Cited on page 198.)

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, pages 997–1003. (Cited on page 198.)

Coletta, V. P. and Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(12):1172–1182. (Cited on page 190.)

Coller, B. D. (2008). An experiment in hands-on learning in engineering mechanics: statics. *International Journal of Engineering Education*, 24(3):545. (Cited on page 65.)

Crocker, L. M. and Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston. (Cited on pages 5, 39, 41, and 70.)

Cronbach, L. J. and Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, 74(1):68–80. (Cited on pages 184 and 190.)

Dankert, J. and Dankert, H. (2013). *Technische Mechanik: Statik, Festigkeitslehre, Kinematik/Kinetik ; mit 128 Übungsaufgaben, zahlreichen Beispielen und weiteren Abbildungen im Internet*. Lehrbuch. Springer Vieweg, Wiesbaden, 7th edition. (Cited on pages 12, 14, 118, and 119.)

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., and Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39):19251–19257. (Cited on pages 2, 4, 20, and 247.)

Deslauriers, L., Schelew, E., and Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, 332(6031):862–864. (Cited on page 2.)

Ding, L. and Liu, X. (2012). Getting started with quantitative methods in physics education research. In *Getting Started in PER*, volume 2. 3rd edition. (Cited on page 4.)

Direnga, J. and Kautz, C. (2019). (Re-)validation of a standardised test instrument in a different national context. In *Research in Engineering Education Symposium 2019 (REES2019)*, Cape Town, South Africa. (Cited on page 70.)

Direnga, J., Presentati, B., Timmermann, D., Brose, A., and Kautz, C. (2015a). Does it stick? - Investigating long-term retention of conceptual knowledge in mechanics instruction. In *Proceedings of the 122nd ASEE Annual Conference & Exposition*. (Cited on page 221.)

Direnga, J., Timmermann, D., Brose, A., and Kautz, C. (2014). A statistical method for assessing teaching effectiveness based on non-identical pre-and post-tests. In *Proceedings of the SEFI 2014 Annual Conference*, Birmingham, UK. (Cited on pages 185, 189, and 198.)

Direnga, J., Timmermann, D., Kieckhäfer, F., Brose, A., and Kautz, C. (2018). The discriminative learning gain: a two-parameter quantification of the difference in learning success between courses. *Australasian Journal of Engineering Education*, 23(2):71–82. (Cited on page 185.)

Direnga, J., Timmermann, D., Kieckhäfer, F., and Kautz, C. (2017). Refining the WLR: Quantifying the difference in learning success between courses. In *Research in Engineering Education Symposium 2017 (REES2017)*, Bogotá, Columbia. (Cited on pages 185 and 198.)

Direnga, J., Timmermann, D., Lund, J., and Kautz, C. (2016). Design and application of self-generated identification codes (SGICs) for matching longitudinal data. In *Proceedings of the 44th SEFI Annual Conference*, Tampere, Finland. (Cited on page 206.)

Direnga, J., Timmermann, D., Presentati, B., Brose, A., and Kautz, C. (2015b). Do students spend more time on difficult questions? Analysis of item response time versus correctness in the SCI/CATS. In *Research in Engineering Education Symposium 2015 (REES2015)*, Dublin, Ireland. (Cited on page 172.)

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2-3):105–225. (Cited on page 2.)

diSessa, A. A. (2014). A history of conceptual change research. In Sawyer, R. K., editor, *The Cambridge Handbook of the Learning Sciences*, Cambridge Handbooks in Psychology, pages 88–108. Cambridge University Press, 2nd edition. (Cited on pages 19 and 27.)

Edström, K. (2012). Student feedback in engineering: a discipline-specific overview and background. In *Enhancing Learning and Teaching Through Student Feedback in Engineering*. Elsevier. (Cited on pages 21 and 22.)

Eller, C. (2015). Grundbegriffe und Axiome der Statik starrer Körper. In *Holzmann/Meyer/Schumpich Technische Mechanik Statik*, pages 13–27. Springer Fachmedien Wiesbaden, Wiesbaden. (Cited on pages 12 and 14.)

Engelhardt, P. V. (2009). An introduction to classical test theory as applied to conceptual multiple-choice tests. In *Getting Started in PER*, volume 2. (Cited on pages 37, 41, 42, 74, 85, 86, and 87.)

Engelhardt, P. V. and Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics*, 72(1):98–115. (Cited on page 70.)

Engqvist, L. (2005). The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. *Animal Behaviour*, 70(4):967–971. (Cited on pages 185 and 191.)

Felder, R. M., Felder, G. N., and Dietz, E. J. (1998). A longitudinal study of engineering student performance and retention. V. Comparisons with traditionally-taught students. *Journal of Engineering Education*, 87(4):469–480. (Cited on page 20.)

Finkelstein, N. D. and Pollock, S. J. (2005). Replicating and understanding successful innovations: Implementing tutorials in introductory physics. *Physical Review Special Topics - Physics Education Research*, 1(1). (Cited on pages 3, 24, 25, 33, 187, 224, 245, and 246.)

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and 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. (Cited on pages 2, 22, 23, and 247.)

Garvin-Doxas, K., Doxas, I., and Klymkowsky, M. W. (2007). Ed's tools: a web-based software toolset for accelerated concept inventory construction. In *Proceedings of the National STEM Assessment Conference*, Washington, D.C. (Cited on page 42.)

Geier, C. (2016). *Analyse eines standardisierten Tests zum Konzeptverständnis im Fach Statik zur Verwendung als Eingangstest*. Unpublished bachelor thesis, Hamburg University of Technology (TUHH), Hamburg. (Cited on pages 73, 76, 183, and 283.)

Girwidz, R., Kurz, G., and Kautz, C. (2003). Zum Verständnis der newtonschen Mechanik bei Studienanfängern–der Test ‚Force Concept Inventory–FCI. In Nordmeier, V., editor, *Didaktik der Physik. Beiträge der Frühjahrstagung der DPG–Augsburg*, Berlin. (Cited on pages 187 and 242.)

Goncher, A. M. and Boles, W. (2019). Enhancing the effectiveness of concept inventories using textual analysis: investigations in an electrical engineering subject. *European Journal of Engineering Education*, 44(1-2):222–233. (Cited on page 41.)

Gross, D., Hauger, W., Schröder, J., and Wall, W. (2011). *Technische Mechanik 1*. Springer-Lehrbuch. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on pages 3, 11, 12, 13, 118, 119, and 121.)

Ha, O., Brown, S., and Pitterson, N. (2017). An exploratory factor analysis of Statics Concept Inventory data from practicing civil engineers. *International Journal of Engineering Education*, 33(1):236–246. (Cited on page 66.)

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. (Cited on pages ii, 2, 4, 5, 23, 24, 45, 184, 187, 188, and 189.)

Hake, R. R. (2002). Assessment of physics teaching methods. In *Proceedings of the UNESCO-ASPEN Workshop on Active Learning in Physics*, Univ. of Peradeniya, Sri Lanka. (Cited on page 184.)

Hake, R. R. (2010). *Should we measure change? Yes*. American Evaluation Association. (Cited on pages 184, 190, and 247.)

Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3):309–333. (Cited on page 40.)

Halloun, I. A. and Hestenes, D. (1985a). Common sense concepts about motion. *American journal of physics*, 53(11):1056–1065. (Cited on page 42.)

Halloun, I. A. and Hestenes, D. (1985b). The initial knowledge state of college physics students. *American journal of Physics*, 53(11):1043–1055. (Cited on page 45.)

Hamburg University of Technology (2019). Module Description: Mechanics I (Statics). https://intranet.tuhh.de/kvvz/vorlesung.php?Lang=en&sg_s=MBBC&mid=889&xtype=s&kvvz=1. (Cited on pages 117, 119, 214, and 216.)

Hammer, D. (1996). More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research. *American Journal of Physics*, 64(10):1316–1325. (Cited on page 19.)

Heller, P. and Huffman, D. (1995). Interpreting the Force Concept Inventory - A reply to Hestenes and Halloun. *The Physics Teacher*, 33(8):503–511. (Cited on page 45.)

Henderson, C. (2002). Common concerns about the Force Concept Inventory. *The Physics Teacher*, 40(9):542–547. (Cited on page 242.)

Henderson, C. and Dancy, M. H. (2007). Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics - Physics Education Research*, 3(2):020102. (Cited on page 246.)

Heron, P. R. (2015). Effect of lecture instruction on student performance on qualitative questions. *Physical Review Special Topics - Physics Education Research*, 11(1):010102. (Cited on page 2.)

Heron, P. R. (2018). Identifying and addressing difficulties: Reflections on the empirical and theoretical basis of an influential approach to improving physics education. In *Getting Started in PER*, volume 2. 4th edition. (Cited on pages 19, 25, 28, 35, and 247.)

Heron, P. R. L., Loverude, M. E., Shaffer, P. S., and McDermott, L. C. (2003). Helping students develop an understanding of Archimedes' principle. II. Development of research-based instructional materials. *American Journal of Physics*, 71(11):1188. (Cited on page 2.)

Hestenes, D. and Halloun, I. (1995). Interpreting the Force Concept Inventory - A response to March 1995 critique by Huffman and Heller. *The Physics Teacher*, 33(8):502–506. (Cited on pages 45 and 46.)

Hestenes, D. and Wells, M. (1992). A Mechanics Baseline Test. *The Physics Teacher*, 30(3):159–166. (Cited on pages ii, 5, 45, 46, and 47.)

Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30(3):141–158. (Cited on pages ii, 4, 42, 43, 44, 45, and 154.)

Hibbeler, R. C. (2004). *Engineering mechanics: statics*. Pearson Prentice Hall, Singapore, 3rd edition. (Cited on pages 118, 119, 120, and 121.)

Huffman, D. and Heller, P. (1995). What does the Force Concept Inventory actually measure? *The Physics Teacher*, 33(3):138–143. (Cited on pages 45 and 46.)

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The journal of wildlife management*, pages 763–772. (Cited on page 198.)

Johri, A. and Olds, B. M., editors (2013). *Cambridge Handbook of Engineering Education Research*. Cambridge University Press, New York. (Cited on page 20.)

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., and Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, 104(4):454–496. (Cited on pages xv, 37, 50, 63, 64, 65, 66, 79, 87, 95, 100, 127, 156, 159, 164, 166, 167, 169, 178, 179, and 291.)

Joyner, S. and Molina, C. (2012). Class time and student learning. Briefing paper, Texas Comprehensive Center at SEDL. (Cited on page 226.)

Kautz, C. (2010). *Tutorien zur Elektrotechnik*. Pearson Studium - Elektrotechnik. Pearson Studium, München. (Cited on page 26.)

Kautz, C. (2014). Verständnisschwierigkeiten und Fehlvorstellungen in Grundlagenfächern des ingenieurwissenschaftlichen Studiums. In Rentschler, M. and Metzger, G., editors, *Perspektiven angewandter Hochschuldidaktik – Studien und Erfahrungsberichte.*, volume 44 of *Report – Beiträge zur Hochschuldidaktik*, pages 81–131. Shaker, Aachen. (Cited on page 19.)

Kautz, C., Brose, A., and Hoffmann, N. (2018). *Tutorien zur Technischen Mechanik: Arbeitsmaterialien für das Lehren und Lernen in den Ingenieurwissenschaften*. Springer Vieweg. (Cited on pages vii, 4, 26, 32, 33, 217, and 244.)

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1):17–24. (Cited on page 87.)

Kidron, Y. and Lindsay, J. (2014). The effects of increased learning time on student academic and nonacademic outcomes: Findings from a meta-analytic review. Technical Report REL 2014–015, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia, Washington, DC. (Cited on page 226.)

Koenig, K. M. and Endorf, R. J. (2003). Study of TA's ability to implement the Tutorials in Introductory Physics and student conceptual understanding. In *Physics Education Research Conference 2003*, volume 720 of *PER Conference*, pages 161–164, Madison, WI. AIP Publishing. (Cited on page 34.)

Koenig, K. M., Endorf, R. J., and Braun, G. A. (2007). Effectiveness of different tutorial recitation teaching methods and its implications for TA training. *Physical Review Special Topics - Physics Education Research*, 3(1):010104. (Cited on pages 34, 237, and 246.)

Kryjevskaia, M., Boudreaux, A., and Heins, D. (2014). Assessing the flexibility of research-based instructional strategies: Implementing Tutorials in Introductory Physics in the lecture environment. *American Journal of Physics*, 82(3):238–250. (Cited on page 34.)

Lindell, R. and Ding, L. (2013). Establishing reliability and validity: An ongoing process. *AIP Conference Proceedings*, 1513(1):27–29. Publisher: American Institute of Physics. (Cited on pages 5, 50, and 69.)

Lindell, R. S., Peak, E., and Foster, T. M. (2006). Are they all created equal? A comparison of different concept inventory development methodologies. In *PERC Proceedings*, volume 883, pages 14–17. AIP. (Cited on pages 37 and 50.)

Litzinger, T., Firetto, C., Passmore, L., Meter, P. V., Higley, K., Masters, C. B., Costanzo, F., Gray, G. L., Turns, S., and Kulikowich, J. (2008). Identifying and remediating deficiencies in problem solving in statics. In *Proceedings of the 2008 Annual Conference & Exposition*, pages 13.680.1–13.680.18, Pittsburgh, Pennsylvania. https://peer.asee.org/3286. (Cited on page 65.)

Loverude, M. E., Heron, P. R. L., and Kautz, C. H. (2010). Identifying and addressing student difficulties with hydrostatic pressure. *American Journal of Physics*, 78(1):75–85. (Cited on page 33.)

Madsen, A., McKagan, S. B., and Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is incon-

sistent, and what factors influence the gap? *Physical Review Special Topics - Physics Education Research*, 9(2):020121. (Cited on page 41.)

Marx, J. D. and Cummings, K. (2007). Normalized change. *American Journal of Physics*, 75(1):87–91. (Cited on pages 6, 184, 188, and 189.)

Mazak, C. M., Font-Santiago, C. B., and Santiago-Román, A. I. (2014). Effects of language on CATS performance. In *Proceedings of the 2014 ASEE Annual Conference & Exposition*, pages 24.462.1–24.462.12. (Cited on page 70.)

Mazur, E. (1997). *Peer instruction: A user's manual*. Prentice Hall, Upper Saddle River, N.J. (Cited on pages 5, 22, 42, and 245.)

McBride, D. L., Murphy, S., Zollman, D. A., Singh, C., Sabella, M., and Rebello, S. (2010). Student understanding of the correlation between hands-on activities and computer visualizations of NMR/MRI. pages 225–228, Portland, (Oregon). (Cited on page 2.)

McCray, R., DeHaan, R. L., and Schuck, J. A., editors (2003). *Improving undergraduate instruction in science, technology, engineering, and mathematics: Report of a workshop*. National Academies Press, Washington, DC. (Cited on page 20.)

McDermott, L. C. (1991). Millikan Lecture 1990: What we teach and what is learned - Closing the gap. *American Journal of Physics*, 59(4):301–315. (Cited on pages 3 and 25.)

McDermott, L. C. (2001). Oersted Medal Lecture 2001: "Physics education research - The key to student learning". *American Journal of Physics*, 69(11):1127. (Cited on pages 2, 4, 22, 29, 209, and 226.)

McDermott, L. C. and Redish, E. F. (1999). Resource letter: PER-1: Physics education research. *American Journal of Physics*, 67(9):755–767. (Cited on page 24.)

McDermott, L. C. and Shaffer, P. S. (1998). *Tutorials in introductory physics*. Prentice Hall. (Cited on pages vii, 2, 25, 33, and 217.)

Meriam, J. L. and Kraige, L. G. (2008). *Engineering mechanics*. J. Wiley, New York, 6th edition. (Cited on pages ii, 1, 11, 15, 16, 119, and 120.)

Minstrell, J. (1981). Arnold Arons. *The Physics Teacher*, 19(8):520–526. (Cited on page 25.)

Montfort, D., Brown, S., and Pollock, D. (2009). An investigation of students' conceptual understanding in related sophomore to graduate-level engineering and mechanics courses. *Journal of Engineering Education*, 98(2):111–129. (Cited on pages 20, 42, and 105.)

Moosbrugger, H. (2012). *Testtheorie und Fragebogenkonstruktion mit 41 Tabellen*. Springer, Berlin [u.a.]. (Cited on pages 37, 86, 87, and 88.)

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., Mzoughi, T., and McCauley, V. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics*, 74(5):449–453. (Cited on pages 91, 125, and 135.)

Moskal, B. M., Reed, T., and Strong, S. A. (2013). Quantitative and mixed methods research. In Johri, A. and Olds, B. M., editors, *Cambridge Handbook of Engineering Education Research*, pages 519–534. Cambridge University Press, New York. (Cited on page 4.)

National Research Council (2001). *Knowing what students know: the science and design of educational assessment*. National Academies Press, Washington, D.C. (Cited on page 39.)

National Science Foundation (2018). NSF award search. https://www.fastlane.nsf.gov/a6/A6Start.htm on March 15, 2019. (Cited on page 5.)

Neter, J., Wasserman, W., and Kutner, M. H. (1985). *Applied linear statistical models*. Irwin, 2nd edition. (Cited on pages 191, 192, 193, 197, 198, 199, 200, and 206.)

Newcomer, J. L. and Steif, P. S. (2007). Gaining insight into student thinking from their explanations of a concept question. In *Proceedings of the 1st International Conference on Research in Engineering Education*. (Cited on pages 169 and 171.)

Newcomer, J. L. and Steif, P. S. (2008). Student thinking about static equilibrium: insights from written explanations to a concept question. *Journal of Engineering Education*, 97(4):481–490. (Cited on pages 3, 61, 62, 65, 169, and 174.)

Nie, Y., Xiao, Y., Fritchman, J. C., Liu, Q., Han, J., Xiong, J., and Bao, L. (2019). Teaching towards knowledge integration in learning force and motion. *International Journal of Science Education*, 41(16):2271–2295. Publisher: Routledge. (Cited on page 24.)

Nieminen, P., Savinainen, A., and Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics - Physics Education Research*, 6(2):020109. (Cited on page 49.)

Nissen, J. M., Talbot, R. M., Nasim Thompson, A., and Van Dusen, B. (2018). Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories. *Physical Review Physics Education Research*, 14(1):010115. (Cited on page 190.)

Novak, G. M., Gavrin, A., Patterson, E., and Christian, W. (1999). *Just-in-time teaching: Blending active learning with web technology*. Prentice Hall series in educational innovation. Prentice Hall. (Cited on pages 214 and 224.)

Ortiz, L. G., Heron, P. R. L., and Shaffer, P. S. (2005). Student understanding of static equilibrium: Predicting and accounting for balancing. *American Journal of Physics*, 73(6):545–553. (Cited on pages 31 and 33.)

Papadopoulos, C., Román, A. I. S., Perez-Vargas, M. J., Portela-Gauthier, G., and Phanord, W. C. (2016). Development of an Alternative Statics Concept Inventory usable as a pretest. In *Proceedings of the 2016 ASEE Annual Conference & Exposition*, New Orleans, Louisiana. (Cited on pages ii and 48.)

Pellegrino, J. W., DiBello, L. V., and Brophy, S. P. (2013). The Science and Design of Assessment in Engineering Education. In Johri, A. and Olds, B. M., editors, *Cambridge Handbook of Engineering Education Research*, pages 571–598. Cambridge University Press, New York. (Cited on pages 50 and 243.)

Piaget, J. (1970). Piaget's theory. In Mussen, P., editor, *Carmichael's manual of child psychology*, volume 1, pages 703–732. Wiley, New York, 3rd edition. (Cited on pages 2 and 27.)

Pollock, S. J. (2009). Longitudinal study of student conceptual understanding in electricity and magnetism. *Physical Review Special Topics-Physics Education Research*, 5(2):020110. (Cited on page 4.)

Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2):211–227. (Cited on pages 2 and 29.)

Revelle, W. (2019). Documentation of the package 'psych' for R. https://cran.r-project.org/web/packages/psych/psych.pdf. (Cited on page 95.)

Richardson, J., Steif, P., Morgan, J., and Dantzler, J. (2003). Development of a concept inventory for strength of materials. In *Proceedings of th 33rd ASEE/IEEE Frontiers in Education Conference*, volume 1, page T3D, Boulder, CO, USA. IEEE. (Cited on page 49.)

Riegler, P., Simon, A., Prochaska, M., Kautz, C., Bierwirth, R., Hagendorf, S., and Kortemeyer, G. (2016). Using Tutorials in Introductory Physics on circuits in a German university course: Observations and experiences. *Physics Education*, 51(6):065014. (Cited on pages 34 and 246.)

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, 77(1):1–15. (Cited on pages 20 and 21.)

Román, A. I. S., Streveler, R., Steif, P., and DiBello, L. (2010a). The development of a Q matrix for the Concept Assessment

Tool for Statics. In *Proceedings of the 2010 Annual Conference & Exposition*, pages 15.1218.1–15.1218.20, Louisville, Kentucky. https://peer.asee.org/16659. (Cited on pages 124, 125, 126, and 304.)

Román, A. I. S., Streveler, R. A., and DiBello, L. (2010b). The development of estimated cognitive attribute profiles for the Concept Assessment Tool for Statics. In *2010 IEEE Frontiers in Education Conference (FIE)*, pages F2G–1–F2G–8. (Cited on pages 126, 166, and 169.)

Rozenblit, L. and Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 92:1–42. (Cited on page 247.)

Sayre, E. C., Franklin, S. V., Dymek, S., Clark, J., and Sun, Y. (2012). Learning, retention, and forgetting of Newton's third law throughout university physics. *Physical Review Special Topics - Physics Education Research*, 8(1):010116. (Cited on page 235.)

Schecker, H. and Gerdes, J. (1999). Messung von Konzeptualisierungsfähigkeit in der Mechanik - Zur Aussagekraft des Force Concept Inventory. *Zeitschrift für Didaktik der Naturwissenschaften*, 5(1):75–89. (Cited on page 242.)

Schunk, D. H. (2012). *Learning theories: An educational perspective*. Pearson, Boston, 6th edition. (Cited on pages 26, 27, and 28.)

Schwartz, M. (2013). Khan Academy: The illusion of understanding. *Online Learning*, 17(4). (Cited on page 247.)

Shaffer, P. S. and McDermott, L. C. (2005). A research-based approach to improving student understanding of the vector nature of kinematical concepts. *American Journal of Physics*, 73(10):921–931. (Cited on pages 33 and 66.)

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2):4–14. (Cited on page 245.)

Singh, C. and University of Pittsburgh Physics Education Research Group (2012). University of Pittsburgh E&M Tutorial Series. https://www.compadre.org/per/items/detail.cfm?ID=12620. (Cited on page 26.)

Smith, K. A., Sheppard, S. D., Johnson, D. W., and Johnson, R. T. (2005). Pedagogies of Eegagement: Classroom-based practices. *Journal of Engineering Education*, 94(1):87–101. (Cited on page 22.)

Steele, K. M., Brunhaver, S. R., and Sheppard, S. D. (2014). Feedback from in-class worksheets and discussion improves performance on the Statics Concept Inventory. *International Journal of Engineering Education*, 30(4):992–999. (Cited on page 65.)

Steif, P. S. (2004). An articulation of the concepts and skills which underlie engineering statics. In *Frontiers in Education, 2004. FIE 2004. 34th Annual*, pages F1F–5–F1F–10. IEEE. (Cited on pages 1, 3, 51, and 54.)

Steif, P. S. and Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education*, 94(4):363–371. (Cited on pages 46, 51, 53, 54, 55, 58, 61, 98, and 116.)

Steif, P. S., Dollar, A., and Dantzler, J. A. (2005). Results from a statics concept inventory and their relationship to other measures of performance in statics. In *Proceedings Frontiers in Education 35th Annual Conference*, pages T3C–5–T3C–10. (Cited on page 98.)

Steif, P. S. and Hansen, M. A. (2006a). Comparisons between performances in a statics concept inventory and course examinations. *International Journal of Engineering Education*, 22(5):1070–1076. (Cited on pages 9, 55, 62, 63, 98, 123, 124, 125, and 126.)

Steif, P. S. and Hansen, M. A. (2006b). Feeding back results from a statics concept inventory to improve instruction. In *Proceedings of the 2006 American Society of Engineering Education Conference and Exposition*. (Cited on page 166.)

Steif, P. S. and Hansen, M. A. (2007). New practices for administering and analyzing the results of concept inventories. *Journal of Engineering Education*, 96(3):205–212. (Cited on pages 9, 51, 55, 56, 60, 63, 73, 93, 98, 179, 181, and 229.)

Steinberg, R. N., Wittmann, M. C., and Redish, E. F. (1997). Mathematical tutorials in introductory physics. In *AIP Conference Proceedings*, volume 399, pages 1075–1092, College Park, Maryland (USA). AIP. (Cited on page 26.)

Streveler, R. A., Brown, S., Herman, G. L., and Montfort, D. (2014). Conceptual change and misconceptions in engineering education. In Johri, A. and Olds, B. M., editors, *Cambridge Handbook of Engineering Education Research*. Cambridge University Press, New York. (Cited on pages 21 and 28.)

Theobald, R. and Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate stem education research. *CBE Life Sciences Education*, 13(1):41–48. (Cited on pages 4, 185, and 206.)

Thornton, R. K., Kuhl, D., Cummings, K., and Marx, J. (2009). Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory. *Physical Review Special Topics-Physics Education Research*, 5(1):010105. (Cited on pages 49 and 196.)

Thornton, R. K. and 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. (Cited on pages ii, 49, and 50.)

Timmermann, D. and Kautz, C. (2015). Multiple-choice questions that test conceptual understanding: A proposal for qualitative two-tier exam questions. In *Proceedings of the 122nd ASEE Annual Conference & Exposition*, Seattle, WA, USA. (Cited on page 41.)

Timmermann, D. and Kautz, C. (2017). *Tutorien zur Informatik*. DOI: 10.15480/882.2407. (Cited on page 26.)

Treagust, D. F. (2012). Diagnostic assessment in science as a means to improving teaching, learning and retention. In *Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference)*. (Cited on page 40.)

Van Heuvelen, A. (1991). Overview, Case Study Physics. *American Journal of Physics*, 59(10):898–907. (Cited on page 2.)

Venters, C., McNair, L., and Paretti, M. (2014). Writing and conceptual knowledge in statics. In *Proceedings of the 2014 FIE Conference*, pages 1907–1914, Madrid, Spain. (Cited on page 65.)

Von Korff, J., Archibeque, B., Gomez, K. A., Heckendorf, T., McKagan, S. B., Sayre, E. C., Schenk, E. W., Shepherd, C., and Sorell, L. (2016). Secondary analysis of teaching methods in introductory physics: A 50 k-student study. *American Journal of Physics*, 84(12):969–974. (Cited on pages 23, 45, and 190.)

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, 4(1):45–69. (Cited on page 2.)

Wittmann, M. C., Steinberg, R. N., Redish, E. F., and University of Maryland Physics Education Research Group (2004). *Activity-based tutorials: Introductory Physics, The Physics Suite*. John Wiley & Sons, Hoboken, NJ. (Cited on page 26.)

Wittmann, M. C., Steinberg, R. N., Redish, E. F., and University of Maryland Physics Education Research Group (2005). *Activity-based tutorials, Volume 2: Modern Physics, The Physics Suite*. John Wiley & Sons, Hoboken, NJ. (Cited on page 26.)

Wosilait, K., Heron, P. R., Shaffer, P. S., and McDermott, L. C. (1998). Development and assessment of a research-based tutorial on light and shadow. *American Journal of Physics*, 66(10):906–913. (Cited on page 33.)