UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Entwicklung eines Kompetenzniveaumodells der Technischen Mechanik für die Studieneingangsphase im Bauingenieurwesen

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Bauwissenschaften,
Technologie und Didaktik der Technik
der Universität Duisburg-Essen,
zur Erlangung des akademischen Grades
Doktor der Philosophie
- Dr. phil. -

genehmigte Dissertation von

Marcel Pelz, M.Sc.
geboren in Hagen

1. Gutachter:	Prof. Dr. phil. Dipl.-Ing. Martin Lang
2. Gutachter:	Prof. Dr. phil. Dipl.-Ing. Stefan Fletcher

Tag der mündlichen Prüfung: 11.03.2024

Essen, 2024

DOI: 10.17185/duepublico/81744
URN: urn:nbn:de:hbz:465-20240320-082141-7

Für meine Familie

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AIC Akaike Information Criterion

α Signifikanzniveau

ALSTER *Akademisches Lernen und Studienerfolg in der Eingangsphase von naturwissenschaftlich-technischen Studiengängen*

ANOVA Varianzanalyse

BIC Bayesian Information Criterion

BMBF *Bundesministerium für Bildung und Forschung*

CAIC Consistent Akaike Information Criterion

cc vollständige Fälle

d Cohen's *d*

df Freiheitsgrade

DFG *Deutsche Forschungsgemeinschaft*

DIF Differential Item Functioning

DZHW *Deutsches Zentrum für Hochschul- und Wissenschaftsforschung*

EAP Expected a Posteriori

F Forschungsfrage

f relative Häufigkeit

FUNDAMENT *Förderung des individuellen Lernerfolgs mittels digitaler Medien im Bauingenieurstudium*

FW Fachwissen

GK Grundkurs

H Hypothese

$\boldsymbol{H_0}$ Nullhypothese

$\boldsymbol{H_1}$ Alternativhypothese

HE Haupterhebung

HZB Hochschulzugangsberechtigung

ICC Itemchararakteristikkurve

iOM interaktive Online-Module

$\boldsymbol{IQA}$ Interquartilabstand

IRT Item-Response-Theorie

ITF Informatives Tutorielles Feedback

KoKoHs *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor*

KoM@ING *Kompetenzmodellierungen und Kompetenzentwicklung, integrierte IRT-basierte und qualitative Studien bezogen auf Mathematik und ihre Verwendung im ingenieurwissenschaftlichen Studium*

KTT Klassische Testtheorie

LK Leistungskurs

$\boldsymbol{M}$ Mittelwert

MAR Missing at random

$\boldsymbol{Max}$ Maximum

MCAR Missing completely at random

$\boldsymbol{Mdn}$ Median

$\boldsymbol{Min}$ Minimum

MF Modellierungsfähigkeit

MG mathematische Grundlagen

MNAR Missing not at random

MZP Messzeitpunkt

N Grundgesamtheit

n Stichprobengröße

NA fehlende Werte

NE Nacherhebung

NG naturwissenschaftliche Grundlagen

NRW *Nordrhein-Westfalen*

OECD *Organisation for Economic Co-operation and Development*

OSA Online-Self-Assessment

OV Online-Vorkurs

p Wahrscheinlichkeit

φ Phi-Koeffizitent

PISA *Programme for International Student Assessment*

PMM *Predictive Mean Matching*

r Pearson Korrelationskoeffizient

r_s Spearman-Korrelationskoeffizient

RF Rechenfähigkeit

RUB *Ruhr-Universität Bochum*

SD Standardabweichung

SEFI *European Society for Engineering Education*

SoSe Sommersemester

TIMSS *Trends in International Mathematics and Science Study*

TM Technische Mechanik

$\mathbf{TM_4}$ Technische Mechanik 2-Items an MZP 4

TS Summenscore

U Mann-Whitney-Test-Statistik

UDE *Universität Duisburg-Essen*

V Cramer's V

WiSe Wintersemester

wMNSQ Weighted-Mean-Square

z Z-Statsitik

Zusammenfassung

Der Bedarf an Ingenieuren steigt seit einigen Jahren, aber die Zahl der ausgebildeten Ingenieure reicht nicht aus, um den Bedarf zu decken. Neue Ingenieure müssen an den Hochschulen ausgebildet werden, um dem steigenden Bedarf gerecht zu werden. Die Zahl der Studienanfänger in den Ingenieurwissenschaften ist jedoch rückläufig. Hinzu kommt eine hohe Studienabbruchquote in den allgemeinen Ingenieurwissenschaften oder speziell im Bauingenieurwesen. Als einer der Hauptfaktoren für den Studienabbruch in den ersten Fachsemestern gelten Leistungsprobleme, die zumeist auf einen Rückgang spezieller fachlicher, auch mathematischer Kenntnisse bei Studienanfängern zurückzuführen sind und zu Problemen in den Grundlagenfächern wie der Technischen Mechanik oder Mathematik führen. Da die Studienanfänger diese Wissenslücken in der Regel nicht selbstständig schließen können, ist es aus Sicht der Hochschulen wünschenswert, diese gefährdeten Studienanfänger zu identifizieren, um geeignete Unterstützungsmaßnahmen konzipieren und anbieten zu können.

Aus diesem Grund wurde im Rahmen dieser Arbeit ein Kompetenzniveaumodell der Technischen Mechanik für die Studieneingangsphase im Bauingenieurwesen entwickelt. Als Grundlage für die Niveaumodellierung dienen die im Rahmen der Studie ALSTER entwickelten Testinstrumente. Diese Testinstrumente wurden von der Studie FUNDAMENT adaptiert und auch zur Testung von Probanden in den ersten beiden Fachsemestern eingesetzt. Die Daten beider Studien, die an zwei Standorten (*Universität Duisburg-Essen* und *Ruhr-Universität Bochum*) erhoben wurden, dienen somit als Ausgangspunkt für diese Arbeit.

In einem ersten Schritt wurde die gesamte Stichprobe beschrieben und auf signifikante Unterschiede zwischen den beiden Studien unter-

sucht. Es zeigt sich, dass sich sowohl die beiden Studien als auch die Standorte signifikant unterscheiden. Diese Unterschiede können jedoch als typische Jahrgangs- bzw. Kohortenunterschiede interpretiert werden, da die Erhebungen zwei Jahre auseinander liegen. Anschließend erfolgte die Aufbereitung der Daten, wobei der Datensatz mittels eines visuellen Verfahrens (Box-Whisker-Diagramm) auf Ausreißer und Extremwerte untersucht wurde. Im Anschluss erfolgte die Überprüfung des Datensatzes auf fehlende Werte, die mit Hilfe des *Predictive Mean Matching*-Verfahren imputiert wurden, der imputierte Datensatz wurde nochmals auf Ausreißer/Extremwerte untersucht. Der aufbereitete Datensatz wurde nach der Item-Response-Theorie skaliert. Zunächst wurde geprüft, ob eine gemeinsame Skalierung der beiden vorliegenden Teildatensätze zulässig ist. Dabei stellte sich heraus, dass dies für die Technische Mechanik möglich ist, die Testinstrumente der Rechenfähigkeit jedoch für die beiden Studien getrennt skaliert werden müssen. Nach der Untersuchung des Datensatzes auf auffällige FIT-Werte wurden Modellvergleiche für die einzelnen Teilbereiche (Technische Mechanik und Rechenfähigkeit) und Messzeitpunkte durchgeführt. Für die Testinstrumente der Technischen Mechanik und Rechenfähigkeit (FUNDAMENT) zeigte das 2pl-Birnbaum-Modell die beste Passung zu den vorliegenden Daten, während für ALSTER (Rechenfähigkeit) das 1pl-Rasch-Modell am besten passte. Die IRT skalierten Daten wurden dann für die Modellierung der Niveaumodelle verwendet, wobei aufgrund der 2pl-Birnbaum-Skalierung die Lösungswahrscheinlichkeit der Itemschwierigkeiten auf 80% gesetzt wurde. Es zeigt sich, dass die Mehrheit der Probanden nur niedrige Niveaus erreicht. Zudem sind diese Probanden in der Regel nicht in der Lage, an späteren Messzeitpunkten ein höheres Niveau zu erreichen, während dies Probanden gelingt, die zu Beginn einem mittleren Niveau zugeordnet werden. Weiterhin wurde deutlich, dass es Prädiktoren in Form von Schulbildung, Migrations- und Bildungshintergrund gibt, die einen signifikanten Einfluss auf das erreichte Niveau zu Beginn des ersten Studiensemesters haben.

Bei der Überprüfung der Forschungsfragen konnten signifikante Unterschiede in den Personenfähigkeiten der beiden Studien nachgewiesen werden. Darüber hinaus besteht ein signifikanter Zusammenhang zwi-

schen dem erreichten Kompetenzniveau und den erzielten Klausurnoten (Technische Mechanik und Mathematik) am Ende der ersten beiden Fachsemester. Probanden mit einem niedrigen Kompetenzniveau können als Risikogruppe bezeichnet werden, da sie die Klausuren tendenziell nicht bestehen. Ebenso konnte gezeigt werden, dass die in FUNDAMENT entwickelten digitalen Elemente für die Studienvorphase (Online-Self-Assessment und Online-Vorkurs) keinen signifikanten Einfluss auf die Klausurnoten haben. Hingegen können die ebenfalls in FUNDAMENT entwickelten interaktive Online-Module als Indikator für das Erreichen eines bestimmten Kompetenzniveaus angesehen werden.

Abstract

The demand for engineers has been increasing for several years, but the number of trained engineers is not sufficient to meet the demand. New engineers need to be trained at universities in order to meet the increasing demand. However, the number of new engineering students is declining. In addition, there is a high drop-out rate in the general engineering sciences or specifically in civil engineering. One of the main factors for students dropping out in the first few semesters is considered to be performance problems, which are mostly due to a decline in specific technical, including mathematical, knowledge among first-year students and lead to problems in basic subjects such as engineering mechanics or mathematics. s first-year students are generally unable to close these knowledge gaps independently, it is desirable from the universities' point of view to identify these vulnerable first-year students in order to be able to design and offer suitable support measures.

For this reason, a competence level model of engineering mechanics for the introductory phase of civil engineering was developed as part of this thesis. The test instruments developed as part of the ALSTER study serve as the basis for the competence level modelling. These test

instruments were adapted from the FUNDAMENT study and also used to test students in the first two semesters. The data from both studies, which were collected at two locations (University of Duisburg-Essen and Ruhr-University Bochum), thus serve as the starting point for this work.

In a first step, the entire sample was described and analysed for significant differences between the two studies. It was found that both the two studies and the locations differed significantly. However, these differences can be interpreted as typical vintage or cohort differences, as the surveys were conducted two years apart. The data was then processed, whereby the data set was analysed for outliers and extreme values using a visual method (box-whisker diagram). The data set was then checked for missing values, which were imputed using the *Predictive Mean Matching* method, and the imputed data set was analysed again for outliers/extreme values. The processed data set was scaled using the Item-Response-Theorie. First, it was checked whether a joint scaling of the two available partial data sets is permissible. It turned out that this was possible for the engineering mechanics, but that the numeracy test instruments had to be scaled separately for the two studies. After analysing the data set for conspicuous FIT values, model comparisons were carried out for the individual sub-areas (engineering mechanics and mathematics) and measurement point. For the test instruments of engineering mechanics and numeracy (FUNDAMENT), the 2pl-Birnbaum model showed the best fit to the available data, while for ALSTER (numeracy) the 1pl-Rasch model fitted best. The IRT scaled data were then used to model the competence level models, whereby the solution probability of the item difficulties was set to 80% due to the 2pl-Birnbaum scaling. It can be seen that the majority of test subjects only achieve low levels. In addition, these test subjects are generally not able to reach a higher level at later measurement points, while test subjects who are assigned to an intermediate level at the beginning are able to do so. Furthermore, it became clear that there are predictors in the form of school education, migration and educational background that have a significant influence on the level achieved at the beginning of the first semester of study.

When examining the research questions, significant differences were found in the personal abilities of the two studies. In addition, there is a significant correlation between the level of competence achieved and the exam grades achieved (engineering mechanics and mathematics) at the end of the first two semesters. Subjects with a low level of competence can be labelled as a risk group, as they tend to fail the exams. It was also shown that the digital elements developed in FUNDAMENT for the preliminary study phase (online self assessment and online preliminary course) have no significant influence on the exam grades. On the other hand, the interactive online modules also developed in FUNDAMENT can be seen as an indicator for the achievement of a certain level of competence.

Kapitel 1

Einleitung

In den letzten Jahren ist der Bedarf an Ingenieuren[1] deutlich gestiegen (VDI Verein Deutscher Ingenieure e.V. & Institut der Deutschen Wirtschaft e.V. [VDI & iW], 2023). Da dieser Bedarf nicht mit den bereits ausgebildeten Ingenieuren gedeckt werden kann, müssen entsprechende Fachkräfte ausgebildet werden. Allerdings sind die Studienanfängerzahlen in den Ingenieurwissenschaften rückläufig (Statistisches Bundesamt [Destatis], 2023). Eine gesonderte Betrachtung der Studienanfänger im Bauingenieurwesen zeigt ein ähnliches Bild: In Nordrhein-Westfalen (NRW) sind die Studienanfängerzahlen im Bauingenieurwesen in den letzten beiden Jahren um rund 11% zurückgegangen (Destatis, 2023).

Neben sinkenden Studienanfängerzahlen, haben die Ingenieurwissenschaften mit hohen Studienabbruchquoten zu kämpfen. Während die Studienabbruchquote in den allgemeinen Ingenieurwissenschaften bei rund 35% liegt, beträgt sie im Bauingenieurwesen sogar 46% (Heublein et al., 2022). Die Faktoren, die zu einem Studienabbruch führen, sind vielfältig. Als einer der entscheidenden Faktoren gelten Leistungsprobleme der Studierenden (Heublein et al., 2017). Diese Leistungsprobleme führen zu mangelndem Studienerfolg, der in den ersten Studiensemestern zum Studienabbruch führt (Henn & Polaczek, 2007). Die genannten Leistungsprobleme treten häufig in Grundlagenfächern wie der Technischen Mechanik (TM) oder der Mathematik auf (Heublein et al.,

[1] Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung männlicher, weiblicher und diverser Sprachformen (m/w/d) verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

2010). Als mögliche Ursache wird ein Rückgang spezieller fachlicher, aber auch mathematischer Kenntnisse bei Studienanfängern angesehen (u.a. Henn & Polaczek, 2007; Heublein & In der Smitten, 2013). Die so vorhandenen Wissenslücken können nur unter großen Schwierigkeiten geschlossen werden (u.a. Dammann & Lang, 2019; Willige et al., 2014).

Aus Sicht der Hochschulen ist es daher wünschenswert, gefährdete Studienanfänger möglichst frühzeitig zu identifizieren. Dies kann u.a durch Kompetenzdiagnostik geschehen, wobei auch Kompetenzniveaumodelle hilfreich sein könnten. Zu diesem Zweck wurden im Rahmen der Studie *Akademisches Lernen und Studienerfolg in der Eingangsphase von naturwissenschaftlich-technischen Studiengängen* (ALSTER) Testinstrumente zur Erfassung der fachlichen Kompetenzen der Technischen Mechanik und der Rechenfähigkeit entwickelt. Diese Testinstrumente könnten Hinweise auf gefährdete Studienanfänger geben, sodass gezielte Fördermaßnahmen eingeleitet werden können. Derartige Unterstützungsangebote wurden in der Studie *Förderung des individuellen Lernerfolgs mittels digitaler Medien im Bauingenieurstudium* (FUNDAMENT) entwickelt. Die in FUNDAMENT entwickelten digitalen Angebote reichen von der Studienvorphase bis zur Studieneingangsphase. Dazu gehören ein Online-Self-Assessment (OSA), ein Online-Vorkurs (OV) und interaktive Online-Module (iOM). Darüber hinaus wurden die Testinstrumente von ALSTER in FUNDAMENT adaptiert.

Ziel dieser Arbeit ist die Entwicklung eines Kompetenzniveaumodells für die Studieneingangsphase im Bauingenieurwesen. Der Schwerpunkt liegt dabei auf der Technischen Mechanik, aber auch die Rechenfähigkeiten werden untersucht. Die entwickelten Niveaumodelle sollen dazu dienen, mögliche Risikogruppen hinsichtlich des Studienabbruchs zu identifizieren. Ausgangspunkt sind die im Rahmen der Studie ALSTER entwickelten Testinstrumente der Technischen Mechanik zur Erfassung des Fachwissens und der Modellierungsfähigkeit. Sowie die ebenfalls in ALSTER entwickelten Testinstrumente zur Erhebung der Rechenfähigkeit. Diese Testinstrumente wurden in der Studie FUNDAMENT adaptiert und eingesetzt. Somit stehen für diese Arbeit Datensätze aus beiden Studien für die ersten beiden Fachsemester des Bauingenieurstudiums zur Verfügung. Die Datenerhebung fand in FUNDAMENT an der

Universität Duisburg-Essen (UDE) statt, während in ALSTER Daten sowohl an der UDE als auch an der *Ruhr-Universität Bochum* (RUB) erhoben wurden. Beide Standorte sind Ruhrgebietsuniversitäten. In der Studieneingangsphase wurden Daten an drei Messzeitpunkten (MZP) erhoben: Zu Beginn des ersten Fachsemesters (MZP 2), nach dem ersten Fachsemester (MZP 3) und am Ende des zweiten Fachsemesters (MZP 4). Diese Daten werden nach der Item-Response-Theorie (IRT) skaliert und anschließend für die Niveaumodellierung verwendet.

Es soll überprüft werden, ob die erhobenen Daten der beiden Studien signifikante Unterschiede aufweisen oder ob FUNDAMENT als Replikationsstudie betrachtet werden kann (F1). Ebenso soll geklärt werden, ob die entwickelten Niveaumodelle in der Lage sind, den Studienerfolg im Sinne von Klausurnoten vorherzusagen (F2). Darüber hinaus werden demographische Variablen untersucht, die als Prädiktoren für das Erreichen eines bestimmten Kompetenzniveaus zu Beginn des Studiums geeignet sind, sodass mögliche Risikogruppen bereits zu Studienbeginn identifiziert und ggf. Fördermaßnahmen erhalten können. Neben den bereits genannten MZP gibt es einen weiteren, der in der Studienvorphase angesiedelt ist, an dem allerdings ausschließlich in FUNDAMENT Daten erhoben wurden. Dieser MZP bezieht sich auf die Nutzung und Wirksamkeit des OSA (F3) und wird unabhängig von den entwickelten Niveaumodellen analysiert. Abschließend werden die in FUNDAMENT entwickelten iOM für die ersten beiden Fachsemester der Technischen Mechanik auf einen signifikanten Zusammenhang mit dem erreichten Kompetenzniveau untersucht (F4).

Dazu wird im folgenden Kapitel (Kap. 2) der Stand der Forschung dargelegt. Das Konstrukt der Kompetenz und Möglichkeiten zu dessen Erfassung werden vorgestellt. Aktuelle Zahlen zur Studiensituation und zum Studienabbruch in Deutschland und NRW werden präsentiert und Möglichkeiten zur Digitalisierung der Hochschullehre aufgezeigt. Die für die Testinstrumente relevanten fachlichen Inhalte der Technischen Mechanik und Mathematik werden erläutert und die beiden Studien (FUNDAMENT und ALSTER) beschrieben. In Kap. 3 werden die Forschungsfragen und Hypothesen formuliert, während in Kap. 4 die notwendigen Methoden der IRT und der Niveaumodellierung aufgeführt

werden. Das Untersuchungsdesign wird in Kap. 5 beschrieben, hier werden neben den Testitems auch das Testheftdesign erläutert, aber auch die Stichprobe empirisch beschrieben. Kap. 6 stellt den größten Teil dieser Arbeit dar. Hier werden die Datenaufbereitung mit Extremwertanalyse und Imputation des Datensatzes sowie die IRT-Skalierung mit Modellpassung und Modellvergleichen vorgestellt. Ebenso wird die Entwicklung der Niveaumodelle dargestellt und die verschiedenen MZP miteinander verglichen, wobei mögliche Prädiktoren für das Erreichen eines Kompetenzniveaus benannt werden. Das Kapitel schließt mit der Beantwortung der in Kap. 3 formulierten Forschungsfragen. Abschließend wird diese Arbeit in Kap. 7 zusammengefasst.

Kapitel 2

Stand der Forschung

Für die Entwicklung eines Kompetenzniveaumodells für die Technische Mechanik im Bauingenieurstudium ist zunächst eine genaue Klärung des verwendeten Kompetenzbegriffs und der entsprechenden Kompetenzdiagnostik notwendig. Dabei wird neben der allgemeinen Kompetenzdiagnostik auch der etablierte Bereich der schulischen Kompetenzdiagnostik beleuchtet und anschließend der Bezug zum hochschulischen Bereich hergestellt. Die Ingenieurwissenschaften leiden unter hohen Studienabbruchquoten (Heublein et al., 2022). Diesem Problem kann mit digitalen Unterstützungsangeboten entgegengewirkt werden – insbesondere in Grundlagenfächern wie Technische Mechanik oder Mathematik – deren fachspezifische Inhalte sowie drei mögliche Varianten von Unterstützungsangeboten in diesem Kapitel skizziert werden. Abschließend werden die beiden Forschungsprojekte vorgestellt, deren Datenbasis die Grundlage für das Kompetenzniveaumodell bildet.

2.1 Kompetenzen

Wissenschaftliche Hypothesen werden auf der Basis empirischer Daten untersucht. Dazu ist es notwendig, theoretische Konstrukte klar zu definieren, da daraus Operationalisierungen abgeleitet werden (Hartig, 2008). Der wissenschaftliche Diskurs hat keine allgemein befriedigende Definition des Kompetenzbegriffs hervorgebracht. Als ›konzeptuelle Inflation‹ kritisiert Weinert (2001) diesen Umstand. Nach Hartig (2008) ist

das Ziel einer allgemeingültigen Definition allerdings auch unrealistisch, da der Begriff eine zu große alltagssprachliche Bedeutungsvielfalt und unterschiedliche Bedeutungen für einzelne wissenschaftliche Kontexte besitzt. Die Definition und Operationalisierung des Kompetenzkonstrukts beeinflusst die Ergebnisse der Hypothesen, daher ist die Formulierung einer expliziten Arbeitsdefinition für das Verständnis des Begriffs notwendig (Hartig, 2008; Hartig & Klieme, 2006).

In einem Bericht für die *Organisation for Economic Co-operation and Development* (OECD) unterscheidet Weinert (1999) sechs verschiedene Varianten des Kompetenzbegriffs (vgl. Hartig & Klieme, 2006; Klieme, 2004). Weinert (1999, 2001) selbst empfiehlt die Beschränkung auf folgende Definition: »Kompetenzen als kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf bestimmte Klassen von Situationen und Anforderungen beziehen. Diese spezifischen Leistungsdispositionen lassen sich auch als Kenntnisse, Fertigkeiten oder Routinen charakterisieren« (Übersetzung von Hartig und Klieme, 2006, S. 129). Diese Definition fokussiert zwei Eingrenzungen: zum einen sind Kompetenzen kontextspezifisch und damit auf einen bestimmten Bereich von Situationen und Anforderungen beschränkt, zum anderen werden nur kognitive Leistungsdispositionen berücksichtigt (Hartig & Klieme, 2006). Aus pragmatischen Gründen nimmt Weinert (2001) die erste Eingrenzung, die zum Ausschluss der allgemeinen intellektuellen Fähigkeit führt, vor (Hartig & Klieme, 2006). Zur Bewältigung spezifischer Anforderungen sind basale kognitive Grundfunktionen erforderlich, die zu einem allen Menschen gemeinsamen Grundbestand gehören und nicht erworben werden müssen bzw. können, da diese elementaren Funktionen keinen großen Veränderungen durch Lernen oder andere äußere Einflüsse unterliegen (Hartig, 2008; Hartig & Klieme, 2006). Ein Sammelbegriff für diese grundlegenden kognitiven Fähigkeiten ist die Intelligenz (Hartig, 2008). Unter Intelligenz versteht man »die Fähigkeit zur Lösung neuer Probleme ohne spezifisches Vorwissen ..., d.h. es werden darunter breit generalisierbare Leistungsdispositionen zusammengefasst« (Hartig, 2008, S. 18).

Zwischen der Kompetenz und der Intelligenz besteht nicht nur eine starke konzeptionelle Ähnlichkeit, sondern es gibt auch inhaltliche

Anknüpfungspunkte (Hartig & Klieme, 2006). Daher ist eine Gegenüberstellung der beiden Begriffe sinnvoll, wobei nach Hartig und Klieme (2006) eine Abgrenzung anhand von drei Kriterien möglich ist: Kontextualisierung, Lernbarkeit und Binnenstruktur.

Bei der Kontextualisierung unterscheiden sich beide Begriffe hinsichtlich der Breite des Kontextes (Hartig & Klieme, 2006). Während bei der Kompetenz die Kontexte relativ spezifisch sind, sind sie bei der Intelligenz generalisierbar (Hartig & Klieme, 2006). Eine Unterscheidung der beiden Begrifflichkeiten ist nicht mehr eindeutig möglich, sofern der Kontext der Kompetenz zu weit gefasst wird (Hartig & Klieme, 2006).

Wie bereits erwähnt, kann die Kompetenz durch Lernen erworben werden (Hartig & Klieme, 2006; Klieme & Leutner, 2006; Weinert, 2014). Die spezifische Kontextualisierung impliziert, dass Erfahrungen in bestimmten Aufgaben oder Situationen gesammelt werden müssen, damit ein Kompetenzerwerb möglich ist (Hartig & Klieme, 2006). Im Gegensatz dazu wird die Intelligenz als stabiles Persönlichkeitsmerkmal verstanden, dass sich weder über die Zeit verändert noch erlernt werden kann (Hartig & Klieme, 2006). Hartig und Klieme (2006) veranschaulichen dies anhand des Nullpunktes einer Skala. Während die Kompetenz einen klassischen Nullpunkt besitzt – die Kompetenz in dem spezifischen Kontext ist nicht vorhanden – kann bei der Intelligenz kein Nullpunkt erreicht werden, da zumindest basale Grundfähigkeiten bei jedem Menschen vorhanden sind (Hartig & Klieme, 2006).

Als Grundlage für die Erklärung individueller Leistungsunterschiede dienen Binnenstrukturen (Hartig & Klieme, 2006). Diese Strukturen werden bei der Betrachtung von Kompetenzen auf die zu bewältigenden Anforderungen zurückgeführt, während die Intelligenz primär auf psychische Prozesse zurückgeführt wird, bei denen die beobachtbaren Intelligenzleistungen als unverzichtbar und wesentlich angesehen werden (Hartig & Klieme, 2006). Obwohl die drei genannten Kriterien eine sinnvolle konzeptionelle Trennung von Kompetenz und Intelligenz ermöglichen und auch Unterschiede in der Testkonstruktion bestehen, lassen sich durchaus Zusammenhänge zwischen der Messung von Kompetenz und Intelligenz feststellen (Hartig & Klieme, 2006).

Die zweite von Weinert (1999) vorgenommene Fokussierung auf

kognitive Leistungsdispositionen impliziert, dass motivationale oder affektive Einflüsse unberücksichtigt bleiben (Hartig & Klieme, 2006). Allerdings revidiert Weinert (2014) diese Eingrenzung teilweise und definiert den in der empirischen Bildungsforschung häufig zur Operationalisierung verwendeten Kompetenzbegriff als

> die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können. (Weinert, 2014, S. 27–28)

Die Bedeutung dieser Definition für die Entwicklung von Bildungsstandards und Kompetenzmodellen wird von Klieme (2004) noch einmal unterstrichen, indem diese Definition als Referenz in Deutschland hervorgehoben wird.

In empirischen Untersuchungsdesigns empfiehlt Weinert (2001) jedoch eine Trennung von kognitiven und motivationalen Einflüssen (vgl. Hartig, 2008; Klieme & Leutner, 2006). Diese Empfehlung wird auch von Hartig (2008) unterstützt, da ein enger gefasstes Konstrukt sowohl die Genauigkeit als auch den Informationsgehalt verbessern kann. Als weiteres Argument führt er an, dass Motivation in der Regel als eine über die Zeit veränderliche Größe betrachtet wird (Hartig, 2008). Der empfohlenen Trennung folgen auch Klieme und Leutner (2006) im Schwerpunktprogramm der Deutschen Forschungsgemeinschaft (DFG) ›Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen‹. Kompetenzen werden dort definiert als »›kontextspezifische kognitive Leistungsdispositionen‹, die sich funktional auf Situationen und Anforderungen in bestimmten ›Domänen‹ beziehen« (Klieme und Leutner, 2006, S. 879). Als »gute Arbeitsgrundlage« (Hartig und Klieme, 2006, S. 129) für Fragestellungen in der Bildungsforschung bewerten Hartig und Klieme (2006) diese Definition. In einer weiteren Ausführung hält Hartig (2008) diese Definition von Kompetenzen für nützlich, um persönliche Lernerfolge zu erfassen und Lernfortschritte zu evaluieren, im Weiteren auch für gut geeignet, um

Bildungsprozesse auf ihre Wirkungen hin zu untersuchen (Hartig, 2008).

Auch in dieser Definition wird wieder die ›Kontextabhängigkeit‹ der Kompetenz betont, sie gewährleistet die Erfüllung von Anforderungen, die durch Aufgaben oder Situationen gestellt werden (Hartig, 2008; Klieme & Leutner, 2006). Hartig (2008) kritisiert, dass die oben genannte Kompetenzdefinition durch die freie Wahl der Kontextdefinition eine gewisse Willkür ermöglicht. Deswegen sollte der Kontext »einerseits hinreichend konkret sein, ... andererseits aber auch nicht zu eng gefasst ..., da sonst einfaches Sachwissen oder isolierte Fertigkeiten unnötigerweise als Kompetenzen etikettiert werden« (Hartig, 2008, S. 21).

Im Rahmen dieser Arbeit wird auf die bereits erläuterte Arbeitsdefinition des Kompetenzkonstrukts nach Klieme und Leutner (2006) zurückgegriffen, wonach die Kompetenz als ›kontextspezifische kognitive Leistungsdispositionen‹ definiert wird.

2.1.1 Kompetenzdiagnostik

Nach der Klärung des Kompetenzbegriffs muss zunächst bestimmt werden, wie Kompetenz diagnostiziert werden kann. In der Regel werden dazu standardisierte Tests (Leistungstests) eingesetzt (Hartig & Klieme, 2006). Diese Leistungstests bestehen aus mehreren Aufgaben, die im Folgenden als ›Items‹ bezeichnet werden. Grundlage dieser Tests sind theoretische Kompetenzmodelle, die sich nach Hartig und Klieme (2006) in Kompetenzstrukturmodelle und Kompetenzniveaumodelle unterteilen lassen (vgl. Dorsch et al., 2017, Stichwort: Kompetenzen).

Um die Dimensionalität von Kompetenzen beschreiben zu können, werden Kompetenzstrukturmodelle verwendet (Dorsch et al., 2017; Fleischer et al., 2013; Hartig & Klieme, 2006; Klieme & Leutner, 2006, Stichwort: Kompetenzmodelle). Die Dimensionen entsprechen in diesem Zusammenhang den Teilkompetenzen der betrachteten Kompetenz und stellen unterscheidbare Aspekte dar, um Kompetenzunterschiede zwischen verschiedenen Probanden sichtbar zu machen (Fleischer et al., 2013). Die Dimensionalitätsanalyse erfolgt auf der Grundlage von Analyseverfahren und Kriterien, die auch bei entsprechenden Fragestellungen zur Intelligenz oder allgemeinen Persönlichkeitsmerkmalen verwendet werden (Hartig & Klieme, 2006). Es sind sowohl explorative

als auch theoriebasierte Ansätze möglich (Dorsch et al., 2017, Stichwort: Kompetenzmodelle). Bei dem explorativen Ansatz werden mit Hilfe der Faktorenanalyse Korrelationen zwischen den eingesetzten Testitems aufgedeckt, entsprechend hohe Korrelationen führen dazu, dass die Items zu einer Dimension zusammengefasst werden und somit das gleiche Merkmal messen (Hartig & Klieme, 2006). Ergibt die Faktorenanalyse hingegen keine Zusammenhänge, so liegen die Items nicht auf derselben Dimension und messen somit unterschiedliche Merkmale (Hartig & Klieme, 2006). Die Dimensionen können – wie in der Intelligenz und Persönlichkeitsforschung – als ›latente Variablen‹ verstanden werden (Hartig & Klieme, 2006), die nicht direkt messbar sind (Dorsch et al., 2017, Stichwort: Variable, latente). Der theoriebasierte Ansatz greift auf theoretische Vorüberlegungen zurück, um latente Variablen und deren zu erwartende Zusammenhangsstruktur zu definieren (Hartig & Klieme, 2006). Mit Hilfe von Strukturgleichungsmodellen kann diese aus der Theorie abgeleitete Zusammenhangsstruktur überprüft werden (Hartig & Klieme, 2006).

Die inhaltliche Einordnung konkreter Kompetenzen, die zuvor in unterschiedlicher Ausprägung empirisch erfasst wurden, erfolgt durch Kompetenzniveaumodelle (Hartig & Klieme, 2006). Dabei handelt es sich um einen Ansatz der »›kriteriumsorientierten Interpretation‹ der quantitativen Leistungswerte« (Hartig und Klieme, 2006, S. 133), die auf einer kontinuierlichen Skala verortet werden (Fleischer et al., 2013; Hartig & Klieme, 2006). Diese Skala wird in sinnvolle Abschnitte – sogenannte Kompetenzniveaus – unterteilt, die wiederum inhaltlich die jeweils erfasste Kompetenz eines jeden Abschnitts beschreiben (Hartig & Klieme, 2006). Die Einordnung der Probanden in die jeweiligen Kompetenzniveaus erfolgt anhand der Schwierigkeit der Items, die gerade noch richtig beantwortet werden (Dorsch et al., 2017, Stichwort: Kompetenzmodelle). Fleischer et al. (2013) betonen, dass durch die Modellierung von Kompetenzniveaus Kompetenzen zwar besser beschrieben und auch kommuniziert werden können, dies aber immer mit einer Vereinfachung und Reduktion von Informationen verbunden ist. Weitere Ausführungen zur Kompetenzniveaumodellierung und eine genauere Erläuterung zur Konstruktion der Kompetenzniveaus finden sich in Kap. 4.2.

Weder Kompetenzstrukturmodelle noch Kompetenzniveaumodelle berücksichtigen die zeitliche Variabilität der Kompetenz (Kap. 2.1). Diese Erlernbarkeit und der zeitliche Verlauf der Kompetenz werden – neben den Eigenschaften der beiden anderen Kompetenzmodelle – vor allem in den Naturwissenschaften in Kompetenzentwicklungsmodellen abgebildet (Hammann, 2004). Einige Arbeiten zu Kompetenzentwicklungsmodelle sind bei Leutner et al. (2017) zu finden. Allerdings sind Kompetenzentwicklungsmodelle bisher nur unzureichend theoretisch und empirisch fundiert (Fleischer et al., 2013). Neumann (2020) bezeichnet die empirische Validierung als wenig trivial. Für längere Zeiträume lassen sich Kompetenzentwicklungen identifizieren, für kurze Zeiträume sind die Ergebnisse jedoch nicht eindeutig interpretierbar (Neumann, 2020). Aus diesem Grund – und da die in der Arbeit verwendeten Testinstrumente den aktuellen Kompetenzstand widerspiegeln – wird im Rahmen dieser Arbeit auf weitergehende Ausführungen und die Verwendung von Kompetenzentwicklungsmodellen verzichtet.

Im Hinblick auf Forschungsarbeiten zur Kompetenzmodellierung in ingenieurwissenschaftlichen Studiengängen liegen bisher nur wenige Forschungsarbeiten vor. Drei Forschungsprojekte der Förderlinie *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor* (KoKoHs) (Johannes Gutenberg-Universität Mainz [JGU], o. J. a) des *Bundesministerium für Bildung und Forschung* (BMBF) bilden in diesem Zusammenhang eine Ausnahme. Während im Projekt *Modellierung und Messung von Kompetenzen der Technischen Mechanik in der Ausbildung von Maschinenbauingenieuren* (KOM-ING) (JGU, o. J. b) die Technische Mechanik im Maschinenbau betrachtet wurde, hat sich das Projekt *Kompetenzmodellierungen und Kompetenzentwicklung, integrierte IRT-basierte und qualitative Studien bezogen auf Mathematik und ihre Verwendung im ingenieurwissenschaftlichen Studium* (KoM@ING) (JGU, o. J. c) der Ingenieurmathematik, Technische Mechanik, Werkstoffkunde und Konstruktionstechnik in den Studiengängen Bauingenieurwesen, Elektrotechnik und Maschinenbau gewidmet. Die Ergebnisse des Projektes KoM@ING in Bezug auf die Technische Mechanik berichtet Dammann (2016) in seiner Dissertation. Das Projekt *Modellierung von Kompetenzen bei Studierenden des Maschinenbaus in den Bereichen*

Konstruktion, Entwurf und Produktionstechnik (MoKoMasch) (JGU, o. J. d) befasst sich mit Studierenden des Maschinenbaus und der Verfahrenstechnik. Da bisher jedoch nur wenige ingenieurwissenschaftliche Studien vorliegen, werden Forschungsarbeiten zu inhaltlich verwandten Themen vorgestellt.

2.1.2 Kompetenzdiagnostik in der schulischen Bildung

Im schulischen Bereich sind regelmäßige Leistungstests an der Tagesordnung. Die *Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland* (KMK) hat 2006 mit der Einführung des ›Bildungsmonitorings‹ eine Möglichkeit geschaffen, Entwicklungen im deutschen Bildungssystem (Schülerkompetenzen) kontinuierlich zu beobachten (Kultusministerkonferenz [KMK], 2016). Dazu sind vier Verfahren bzw. Instrumente vorgesehen (KMK, 2016):

1. Teilnahme an internationalen Schulleistungsstudien
2. Überprüfung bzw. Umsetzung von Bildungsstandards für die Primarstufe, die Sekundarstufe I und die Allgemeine Hochschulreife
3. Verfahren zur Qualitätssicherung auf Ebene der Schulen
4. Gemeinsame Bildungsberichterstattung von Bund und Ländern (KMK, 2016, S. 6)

Bei den ersten drei Punkten gibt es inhaltliche Anknüpfungspunkte zur Technischen Mechanik, da u.a. die mathematische, aber auch die naturwissenschaftliche Grundbildung der Schüler im Mittelpunkt steht. Für diese Arbeit weniger relevante Themen wie die Bildungsberichterstattung oder auch die Grundschul-Lese-Untersuchungen (Progress in International Reading Literacy Study [PIRLS]/Internationale Grundschul-Lese-Untersuchung [IGLU]), werden nicht weiter ausgeführt. Relevant sind hingegen die internationalen Schulleistungsstudien *Trends in International Mathematics and Science Study* (TIMSS) und *Programme for International Student Assessment* (PISA).

Die von der *International Association for the Evaluation of Educational Achievement* (IEA) durchgeführte TIMSS-Studie wird seit 1995 alle vier Jahre – 1995 und 1999 unter dem Namen *Third International Mathematics and Science Study* – durchgeführt und ermöglicht einen internationalen Vergleich von Schülerleistungen (Kompetenzen) in der mathematisch-naturwissenschaftlichen Grundbildung (vgl. Baumert et al., 1999; Köller et al., 2001; TIMSS & PIRLS International Study Center [TIMSS & PIRLS ISC], o. J.). Im Mittelpunkt der Untersuchung stehen zwei Populationen: 9-Jährige in der Grundschule (Jahrgangsstufen 3 und 4) und 13-Jährige in der Sekundarstufe I (Jahrgangsstufen 8 und 9) (Baumert et al., 1999). Darüber hinaus wurde 1995 in Deutschland einmalig eine dritte Population von Schülern und Auszubildenden untersucht, »die sich im letzten Jahr einer vollzeitlichen Ausbildung in der Sekundarstufe II befinden« (Baumert et al., 1999, S. 23). Diese Population wird inzwischen als eigenständige Studie unter dem Titel *TIMSS Advanced* in unregelmäßigen Abständen (1995, 2008, 2015) untersucht (TIMSS & PIRLS ISC, o. J.). Deutschland hat jedoch nur in der TIMSS-Studie 1995 alle drei Populationen untersucht (TIMSS & PIRLS ISC, o. J.). An den TIMSS-Studien 1999 und 2003 hat Deutschland nicht teilgenommen und seit 2007 wird nur noch die erste Population in der Jahrgangsstufe 4 untersucht (TIMSS & PIRLS ISC, o. J.). Nachfolgend wird dennoch die Untersuchung der dritten Population – die altersmäßig den Studierenden am nächsten kommt – im Rahmen der TIMSS-Studie 1995 in Deutschland kurz dargestellt. In der ersten von insgesamt drei Teiluntersuchungen wurde die mathematisch-naturwissenschaftliche Grundbildung (gymnasiale Oberstufe und berufliche Schulen) untersucht (Baumert et al., 1999). Dabei wird dem Konzept der *mathematics and science literacy* gefolgt und entsprechend geprüft, ob ein Schüler in der Lage ist, grundlegende mathematische bzw. naturwissenschaftlich-technische Zusammenhänge zu verstehen, die eine Teilhabe am gesellschaftlichen Leben ermöglichen (Köller et al., 2001). Die beiden anderen Teiluntersuchungen fokussieren dagegen ausschließlich die gymnasiale Oberstufe (sowohl Grundkurs [GK] als auch Leistungskurs [LK]), in der die voruniversitären Fachleistungen der Mathematik und Physik erhoben werden (Baumert et al.,

1999). Die Itementwicklung erfolgt mit dem Ziel, durch spezifischen curricularen Kontext die Schülerleistung zu erklären (Köller et al., 2001). Das Multi-Matrix Sampling wurde als Testheftdesign verwendet, die Daten aller drei Teiluntersuchungen wurden nach der probailistischen Testtheorie IRT (siehe Kap. 4.1) skaliert und anschließend in ein (post-hoc) Kompetenzniveaumodell überführt (Baumert et al., 1999) – Vorgehen nach Beaton und Allen (1992) (Kap. 4.2.2).

Für die gymnasiale Oberstufe konnten für die Fachleistungen der Mathematik in der TIMSS-Studie folgende inhaltliche Kompetenzniveaus gebildet werden:

1. Ausführung von mathematischen Routinen
2. Anwendung einfacher Konzepte und Regeln der Oberstufenmathematik
3. selbständiges Anwenden von Lerninhalten der Oberstufe
4. Lösen mathematischer Probleme auf Oberstufenniveau (Baumert et al., 1999).

Die zweite internationale Studie zu den Kompetenzen von Schülern ist die von der OECD initiierte PISA-Studie. Seit dem Jahr 2000 werden alle drei Jahre 15-Jährige Schüler in den Bereichen Lesen, Mathematik und Naturwissenschaften sowie einem weiteren, wechselndem Schwerpunkt – bspw. ›Problemlösekompetenz‹ oder ›Lernen in der digitalen Welt‹(Technische Universität München [TUM], o. J. a) – untersucht (KMK, 2016; Weis & Reiss, 2019). Für die PISA-Studie 2022 wurden noch keine Ergebnisse publiziert, sodass der vorangegangene Durchgang 2018 der jüngste ist. Während zu Beginn der Erhebungen im Jahr 2000 nur 32 Staaten an der Studie teilnahmen, sind es 2018 bereits 79 (Weis & Reiss, 2019). In jedem Durchgang wird einer der drei Kompetenzbereiche als Hauptdomäne stärker in den Fokus gerückt (Weis & Reiss, 2019), während 2018 die Lesekompetenz als Hauptdomäne untersucht wurde, stehen 2022 die Mathematik und 2025 die Naturwissenschaften im Mittelpunkt (Organisation for Economic Co-operation and Development Organisation [OECD], 2023). Hauptdomänen werden

mit besonders vielen Items erhoben und die theoretischen Rahmenkonzepte werden überarbeitet (Weis & Reiss, 2019). Nebendomänen stehen dagegen nicht im Fokus und werden mit weniger Items erfasst, die als Ankeritems (in PISA als ›Trendaufgaben‹ bezeichnet) bereits in den vorherigen Durchgängen eingesetzt wurden (Weis & Reiss, 2019). Ankeritems eignen sich dazu, Items auf einer gemeinsamen Skala zu positionieren, um z.B. die Leistungen an verschiedenen Messzeitpunkten (MZP), hier den verschiedenen PISA-Studien, miteinander vergleichen zu können (Rauch & Hartig, 2020). Wie bei der TIMSS-Studie, steht auch bei PISA der *literacy*-Ansatz im Vordergrund, allerdings tritt bei der Itementwicklung – im Gegensatz zur TIMSS-Studie – die curriculare Validität in den Hintergrund, stattdessen werden Grundkompetenzen in verschiedenen Anwendungsszenarien abgebildet (Baumert et al., 2001). Auch in PISA 2025 bilden die Naturwissenschaften wieder die Hauptdomäne und stützen sich inhaltlich erneut auf die Konzepte der Physik, Chemie, Biologie und Geowissenschaften (Baumert et al., 2001; TUM, o. J. b). Neben dem naturwissenschaftlichen Wissen und Kompetenzen wird auch die naturwissenschaftliche Identität, die für die Förderung der Handlungskompetenz wichtig ist, als hoch relevant für Schüler angesehen (TUM, o. J. b). Als naturwissenschaftliches Wissen wird hier das fachliche, prozedurale und epistemische Wissen verstanden (TUM, o. J. b; vgl. OECD, 2023). Als Anwendungskontexte – mit persönlichem, sozialem oder globalem Bezug – werden Gesundheit, natürliche Ressourcen, Umwelt, Gefahren und Grenzen von Naturwissenschaft und Technik genannt (TUM, o. J. b; vgl. OECD, 2023). Sie sind letztlich in den Kompetenzen ›Phänomene naturwissenschaftlich erklären‹, ›Entwürfe für wissenschaftliche Untersuchungen erstellen und bewerten sowie Daten und Beweise kritisch zu interpretieren‹ und ›wissenschaftliche Informationen recherchieren, bewerten und für die Entscheidungsfindung sowie das Handeln nutzen‹ zu verorten (TUM, o. J. b; vgl. OECD, 2023). Analog zur TIMSS-Studie wurde das Testheftdesign nach dem Verfahren des Multi-Matrix Sampling erstellt und IRT skaliert, anschließend erfolgte die Modellierung eines (post-hoc) Kompetenzmodells (Baumert et al., 2001). Das naturwissenschaftliche Kompetenzmodell von PISA 2018 besteht aus sechs Kompetenzniveaus (Kompetenzniveau 1 ist in 1a und

1b unterteilt, um den unteren Leistungsbereich ausreichend zu differenzieren) (Schiepe-Tiska et al., 2019). In den höheren Kompetenzniveaus werden die Anforderungen an die Schüler entsprechend komplexer: Während Kompetenzniveau 2 das zu erreichende Mindestniveau darstellt, verfügen Schüler der Kompetenzniveaus 5 und 6 über eine ausgeprägte naturwissenschaftliche Kompetenz (Schiepe-Tiska et al., 2019). Die für PISA 2025 vorgesehenen Kompetenzniveaus in den Naturwissenschaften sind in OECD (2023, S. 45) aufgeführt.

Das zweite Element des Bildungsmonitorings, die Bildungsstandards, ermöglichen »einen kompetenzorientierten Unterricht und eine gezielte individuelle Förderung aller Schülerinnen und Schüler ... [sowie] schulische Anforderungen an Schülerinnen und Schüler transparenter, Bildungssysteme durchlässiger und Abschlüsse vergleichbarer zu gestalten« (KMK, 2016, S. 9). Die Bildungsstandards gelten bundesweit im Primarbereich in der Jahrgangsstufe 4 (Deutsch und Mathematik), für die Hauptschule in der Jahrgangsstufe 9 (Deutsch, Mathematik und erste Fremdsprache), für den mittleren Schulabschluss in der Jahrgangsstufe 10 (Deutsch, Mathematik, erste Fremdsprache, Biologie, Chemie und Physik) sowie für die Sekundarstufe II (Deutsch, Mathematik und fortgeführte Fremdsprache) (KMK, 2016). Für den Bereich der Naturwissenschaften in der Sekundarstufe I wurden im Rahmen des Projekts ›Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I‹ (ESNaS) Items entwickelt, die eine Operationalisierung der Kompetenzen der Bildungsstandards in den Naturwissenschaften ermöglichen (Kauertz et al., 2010). Im Unterschied zu den beiden beschriebenen internationalen Vergleichsstudien wird bei ESNaS das Kompetenzmodell nicht post-hoc entwickelt, sondern die Items werden vorab bestimmten Kompetenzausprägungen eines zuvor entwickelten Kompetenzmodells zugeordnet (Kauertz et al., 2010). Das ESNaS-Kompetenzmodell ist als dreidimensionales Modell mit folgenden Dimensionen zu verstehen:

- Komplexität (mit den Niveauabstufung: 1 Fakt, 2 Fakten, 1 Zusammenhang, 2 Zusammenhänge, übergeordnetes Konzept)
- kognitive Prozesse (reproduzieren, selegieren, organisieren, integrieren)

- Kompetenzbereiche (Umgang mit Fachwissen, Erkenntnisgewinnung, Kommunikation, Bewertung) (Kauertz et al., 2010).

Der große Vorteil des Modells liegt in seiner Anwendbarkeit auf alle Fächer, die unter dem Oberbegriff ›Naturwissenschaften‹ zusammengefasst werden (Kauertz et al., 2010). Weitere Details zu ESNaS finden sich in Kauertz et al. (2010), empirische Validierungsergebnisse sind u.a. in V. Fischer (2019), Gehlen (2016), Hostenbach und Walpuski (2013), Kobow (2015), Ropohl (2010), Schoppmeier (2013) und Walpuski et al. (2010) niedergeschrieben.

Die Bildungsstandards haben auch zu zentralen Ländervergleichsstudien geführt, die im Primarbereich alle fünf Jahre (Jahrgangsstufe 4 – Deutsch und Mathematik) und in der Sekundarstufe I alle drei Jahre (Jahrgangsstufe 9 – im Wechsel Deutsch und die erste Fremdsprache bzw. Mathematik und Naturwissenschaften) durchgeführt werden (KMK, 2016). Während die Ländervergleichsstudien und die internationalen Schulleistungsstudien auf repräsentativen Stichproben beruhen, wurde im Rahmen des Bildungsmonitorings mit den ›Verfahren zur Qualitätssicherung auf Ebene der Schulen‹ ein weiteres Instrument geschaffen, das landesweite Untersuchungen auf Schul- oder sogar Klassenebene ermöglicht (KMK, 2016). Diese Vergleichsarbeiten (VERA) werden in der Primarstufe (Jahrgangsstufe 3) und in der Sekundarstufe I (Jahrgangsstufe 8) angeboten (KMK, 2016). In NRW werden VERA in der Jahrgangsstufe 3 in den Fächern Deutsch und Mathematik und in der Jahrgangsstufe 8 in den Fächern Deutsch, der ersten Fremdsprache (Englisch oder Französisch) und Mathematik verpflichtend durchgeführt (Ministeriums für Schule und Bildung NRW [MSB NRW], 2021). VERA sind ausschließlich als Diagnoseinstrument zu verstehen und dürfen nicht zur Benotung einzelner Schüler verwendet werden (MSB NRW, 2021).

Für den schulischen Bereich liegen somit bereits vielfältige Instrumente zur kontinuierlichen Kompetenzdiagnostik vor. Im folgenden Kapitel soll geklärt werden ob entsprechende Instrumente auch für den Hochschulbereich bereits existieren.

2.1.3 Kompetenzdiagnostik im Hochschulsektor

Studien zur Kompetenzdiagnostik im nationalen Hochschulbereich sind zu Beginn der 2010er Jahre rar, insbesondere fehlt es an empirischen Forschungsarbeiten (Zlatkin-Troitschanskaia & Kuhn, 2010). Die wenigen kompetenzdiagnostischen Forschungsansätze genügen nicht den hohen Ansprüchen, wie sie z.B. im schulischen Bereich national und international etabliert sind (z.B. PISA) (Zlatkin-Troitschanskaia & Kuhn, 2010). Zlatkin-Troitschanskaia und Kuhn (2010) weisen zudem darauf hin, dass Leistungsmessungen in Deutschland vor allem die Studierfähigkeit prognostizieren und nicht die im Studium vermittelten Fähigkeiten und Wissensinhalte berücksichtigen. Lediglich Studien im Bereich der Lehrerbildung werden in der systematischen Zusammenfassung von Zlatkin-Troitschanskaia und Kuhn (2010) hinsichtlich der objektiven Kompetenzmessung als »brauchbar« bewertet. Diese Studien zeichnen sich dadurch aus, dass »das fachliche, fachdidaktische und pädagogische Wissen der Zielgruppe auf der Basis eines theoretisch entwickelten fachspezifischen Kompetenzmodells *objektiv*« (Zlatkin-Troitschanskaia und Kuhn, 2010, S. 4) erfasst wird. Während sich die Projekte *COCATIV* und *COCATIV-R* (Baumert & Kunter, 2006; Brunner et al., 2006) auf das Fach Mathematik konzentrieren (Lehrkräfte und Referendare), umfasst *TEDS-LT* (Blömeke et al., 2011) die Fächer Mathematik, Deutsch und Englisch (Lehramtsstudierende und Referendare) (Zlatkin-Troitschanskaia & Kuhn, 2010).

Fortschritte im Bereich der nationalen Kompetenzdiagnostik konnten durch die Förderinitiative ›Hochschulforschung‹ des BMBF, insbesondere mit der 2010 ausgeschriebenen Förderlinie KoKoHs, erzielt werden. Die Ziele der Förderlinie sind:

- die Potenziale der nationalen Hochschulsysteme vor dem Hintergrund des zunehmenden internationalen Wettbewerbsdrucks zu stärken
- die Anbindung an die internationale universitäre Kompetenzforschung sicherstellen
- die Grundlage für nachweislich fundierte organisatorische und individuelle Maßnahmen zur Bewertung von Kompetenzerwerb und

-entwicklung zu schaffen, deren Wirksamkeit überprüft und verbessert werden kann (Blömeke & Zlatkin-Troitschanskaia, 2013).

Blömeke und Zlatkin-Troitschanskaia (2013) nennen hinsichtlich des methodischen Vorgehens psychometrische Modelle, die »neben personenbezogenen (latenten) Merkmalen auch den situativen Anforderungsbezug ... erfassen« (Blömeke und Zlatkin-Troitschanskaia, 2013, S. 9), wofür sich insbesondere die IRT eignet.

Alle Projekte von KoKoHs sind so konzipiert, dass die entwickelten Kompetenzmodelle nachhaltig in den Hochschulen verankert werden können (Blömeke & Zlatkin-Troitschanskaia, 2013). Inhaltlich umfassen die Vorhaben zum einen allgemeine Kompetenzen wie Forschungskompetenz oder Selbstregulation, zum anderen fachspezifische Kompetenzen aus den Erziehungswissenschaften, der MINT-Lehrerbildung, den Wirtschafts- und Sozialwissenschaften sowie den Ingenieurwissenschaften, die mit drei Vorhaben vertreten sind (Blömeke & Zlatkin-Troitschanskaia, 2013). In diesen drei Projekten steht die »Modellierung, Erfassung und Vermittlung von Kompetenzen zur Konstruktion (informations-) technischer Systeme« (Blömeke und Zlatkin-Troitschanskaia, 2013, S. 6) im Vordergrund.

Trotz der genannten Bemühungen »mangelt es im deutschsprachigen Raum ... an validierten Testinstrumenten für eine Eingangs- sowie Lernverlaufsdiagnostik im Hochschulbereich« (Köller et al., 2020, S. 77). Dies hat zu einer Fortsetzung der Förderlinie KoKoHs geführt, die 2015 vom BMBF unter dem Titel *Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen (KoKoHs)* ausgeschrieben wurde und somit die Validierung der entwickelten Kompetenzmodelle und Erhebungsinstrumente entsprechend thematisiert.

In den beiden Förderphasen von 2011 bis 2020 wurden in rund 40 Teilprojekten theoretisch-konzeptionelle Kompetenzmodelle und entsprechende Testinstrumente in verschiedenen Formaten konzipiert und validiert, um Kompetenzen von Studierenden valide und reliabel erfassen zu können (Zlatkin-Troitschanskaia et al., 2020). Sie können drei verschiedene Arten von Konstrukten messen:

- Fähigkeiten, die für den Erfolg der Studierenden in zukünftigen beruflichen Situationen entscheidend sind, auf Wissen basieren und sowohl emotionale als auch motivationale Komponenten beinhalten
- situative Fertigkeiten und Fähigkeiten, die im Studium erlangt wurden
- Leistung der Studierenden unter realen Bedingungen (Zlatkin-Troitschanskaia et al., 2020).

Eine Übersicht über die in den Teilprojekten entwickelten Testinstrumente findet sich in (Zlatkin-Troitschanskaia et al., 2020).

Die Ergebnisse der beiden KoKoHs Förderlinien sind im Metaprojekt KoKoHs-Map von Zlatkin-Troitschanskaia et al. (2021) zusammengefasst. Generell können die »objektiven und zuverlässigen Assessments ... [den] Kompetenzstand von Studierenden entsprechend der aktuellen beruflichen und gesellschaftlichen Anforderungen ... in den untersuchten Domänen valide erfass[en]« (Zlatkin-Troitschanskaia et al., 2021, S. 61). Dies stellt eine deutliche Verbesserung der Kompetenzdiagnostik im Vergleich zu den frühen der 2010er Jahren dar. Die Einteilung aller Projekte in Cluster (Fachkompetenzen, pädagogische Kompetenzen und generische Kompetenzen) zeigt in der Metaanalyse, dass die Kompetenzniveaus der Studierenden in den Projekten des Clusters Fachkompetenzen höher sind als in den beiden anderen Clustern (Zlatkin-Troitschanskaia et al., 2021). Mittelstarke Effekte zeigen sich zwischen Fachkompetenzen und generischen Kompetenzen (Zlatkin-Troitschanskaia et al., 2021). Zlatkin-Troitschanskaia et al. (2021) merken an, dass dieser Befund unterstreicht, dass generische Kompetenzen im Studium meist nicht oder weniger stark gefördert werden als Fachkompetenzen. Am Ende des Studiums wurden zwar bessere Ergebnisse registriert als zu Beginn oder während des Studiums, allerdings unterscheiden sich die Kompetenzniveaus zu Studienbeginn und im Studienverlauf nicht signifikant voneinander (Zlatkin-Troitschanskaia et al., 2021). Dies kann jedoch auch an den eingesetzten Eingangstests liegen, die möglicherweise nicht mit den verwendeten Testinstrumenten vergleichbar sind (Zlatkin-Troitschanskaia et al., 2021). Insgesamt konnten über den gesamten

Erhebungszeitraum signifikante Kompetenzzuwächse beobachtet werden (Zlatkin-Troitschanskaia et al., 2021). Dagegen gelang es den Studierenden nur selten, die eigenen Defizite in den Studieneingangskompetenzen im Laufe des Studiums zu reduzieren, worin die Autoren einen Hinweis auf den sogenannten Matthäus-Effekt sehen (Zlatkin-Troitschanskaia et al., 2021).

Zlatkin-Troitschanskaia et al. (2021) heben explizit hervor, dass die Kompetenzdiagnostik als ergänzendes systematisches Diagnoseinstrument zur zielgerichteten Gestaltung von Lehr-Lern-Maßnahmen und damit zur Verbesserung des Studienerfolg beitragen kann. Hervorgehoben wird auch der Leistungsstand der Studierenden zu Beginn des Studiums, der als entscheidende Grundlage für ein erfolgreiches Studium anzusehen ist (Zlatkin-Troitschanskaia et al., 2021).

2.2 Studium der Ingenieurwissenschaften

Nachdem sich ein Studium an einer deutschen Hochschule in den letzten Jahren immer größeren Beliebtheit erfreute, ist die Zahl der Studierenden im Wintersemester (WiSe) 2022/2023 erstmals seit 15 Jahren zurückgegangen (Destatis, 2023). Mit 2.944.145 Studierenden liegt die Zahl um 1.8% unter dem Vorjahreswert, in NRW ist ein Rückgang um 1.9% zu verzeichnen (Destatis, 2023). Der Rückgang in NRW ist der zweite in Folge, bereits vom WiSe 2020/2021 zum WiSe 2021/2022 war ein Rückgang von 1.9% zu verbuchen. Rückläufige Studierendenzahlen sind in NRW keine Seltenheit, die Schwankungen nach unten liegen aber in der Regel unter 1% (Destatis, 2023).

Während die Zahl der Studierenden gesunken ist, ist die Zahl der Studienanfänger - bezogen auf das Studienjahr, also Sommersemester (SoSe) und das folgende WiSe, leicht gestiegen. Als Studienanfänger werden im Rahmen dieser Arbeit Erstsemesterstudierende verstanden, die sich entweder im ersten Hochschulsemester oder im ersten Fachsemester befinden. Nachdem die Studienanfängerzahlen in den letzten vier Jahren bundesweit rückläufig waren, ist 2023 ein Anstieg um 0.6% zu verzeichnen (Destatis, 2023). Sinkende Studienanfängerzahlen sind in NRW seit dem WiSe 2013/2014 zu beobachten – die Ausnahme bildet das

WiSe 2016/2017 mit einem Anstieg um 0.5% – die Studienanfängerzahlen sind seit diesem Zeitpunkt von 111.062 auf 89.326, also um 19.6% gesunken (Destatis, 2023).

Eine genauere Betrachtung der Studienanfänger in den Ingenieurwissenschaften zeigt, dass die Zahlen rückläufig sind, bis zum WiSe 2022/2023 ist ein Rückgang um 1.3% auf 210.690 Studierende zu verzeichnen (Destatis, 2023). Da sich die im Rahmen dieser Arbeit verwendeten Studien (FUNDAMENT und ALSTER) auf das Studienfach Bauingenieurwesen konzentrieren, wird der Fokus auf dieses Studienfach gerichtet.

Bundesweit wurde im WiSe 2011/2012 mit 11.125 Studierenden der Höchstwert an Studienanfängern im Bauingenieurwesen erreicht, der jedoch bereits im darauf folgenden Semester mit 9.745 Studierenden deutlich unterschritten wurde (Destatis, 2023). Seitdem schwankt der Wert zwischen 8.859 und 9.963, aktuell haben im WiSe 2022/2023 insgesamt 8.998 Studienanfänger ein Studium des Bauingenieurwesens begonnen (Destatis, 2023). Die isolierte Betrachtung der Studienanfänger im Bauingenieurwesen in NRW zeigt ein ähnliches Bild. Auch hier liegt der Spitzenwert im WiSe 2011/2012 mit 2.784 Studierenden, ebenfalls mit einem starken Einbruch im darauf folgenden WiSe (2.558 Studierende) (Destatis, 2023). Seitdem schwanken die Studienanfängerzahlen zwischen 2.090 und 2.558, wobei in den letzten beiden Jahren ein Rückgang um 11.4% auf die aktuelle Anzahl von 2.233 Studierende zu verzeichnen ist (Destatis, 2023).

Der Bedarf an Ingenieuren ist in den letzten Jahren deutlich gestiegen, die bisher ausgebildeten Fachkräfte reichen jedoch nicht aus, um den Bedarf zu decken (VDI & iW, 2023). Der Fachkräftemangel ist nach wie vor ein akutes Problem, wie die Engpasskennziffer für das erste Quartal 2023 in Deutschland zeigt. Die Engpasskennziffer setzt die Arbeitskräftenachfrage in Relation zum Arbeitskräfteangebot und gibt an, wie viele offene Stellen je 100 Arbeitslose vorhanden sind (VDI & iW, 2023). Im ersten Quartal 2023 liegt die Engpasskennziffer für alle Ingenieur- und Informatikberufe bei 456, d.h. auf 100 arbeitslose Ingenieure kommen 456 offene Stellen (in NRW bei 337) (VDI & iW, 2023). Ingenieurberufe im Bauwesen/Vermessungswesen/Gebäudetechnik und

Architektur haben eine Engpasskennziffer von 568 (in NRW bei 497), während Ingenieurberufe im Maschinen- und Fahrzeugbau einen Wert von 435 (in NRW bei 279) erreichen (VDI & iW, 2023). Angesichts der insgesamt rückläufigen Studienanfängerzahlen ist es daher umso wichtiger, dass die Studierenden ihr Studium auch erfolgreich abschließen, um dem Fachkräftemangel entgegenzuwirken.

2.2.1 Studienabbruch

Studienerfolgs- bzw. Studienabbruchquoten dienen als Indikatoren für die Effektivität der akademischen Bildung (Fellenberg & Hannover, 2006). Die Definition des Studienabbruchs ist allerdings nicht trivial, da eine Differenzierung der im Hochschulsystem auftretenden Fluktuationen und Bewegungen berücksichtigt werden muss (Heublein & Wolter, 2011). Dazu gehören nach Heublein und Wolter (2011) »Studienunterbrechung, Studiengangs- oder Fachwechsel, Hochschulwechsel, Hochschulartwechsel oder Wechsel ins Ausland« (Heublein und Wolter, 2011, S. 216). Wird zu diesen Punkten auch noch der Studienabbruch hinzugenommen, spricht man von dem sogenannten ›Schwund‹ bzw. der entsprechenden ›Schwundquote‹ (Heublein & Wolter, 2011). Auf Hochschulebene kann zumeist nur der Schwund ermittelt werden, da der weitere Bildungsverlauf der exmatrikulierten Studierenden in der Regel unbekannt ist (Heublein & Wolter, 2011). Eine Sonderform des Schwunds ist der Studienabbruch, der wiederum nach Bachelor- und Masterstudiengängen unterteilt werden muss. Die Studienabbruchquote für Bachelorstudiengänge gibt den Anteil der Studienanfänger eines Jahrgangs an, die ihr Erststudium nicht erfolgreich abgeschlossen haben und zu keinem späteren Zeitpunkt wieder aufnehmen (Heublein & Wolter, 2011), die Studienerfolgsquote beschreibt entsprechend den verbleibenden Anteil der Studienanfänger, die ihr Erststudium erfolgreich abgeschlossen haben. Heublein et al. (2022) definieren Studienabbrecher (Bachelor) als »Personen, die durch Immatrikulation ein Erststudium an einer deutschen Hochschule aufgenommen ... aber das deutsche Hochschulsystem ohne (ersten) Abschluss verlassen« (Heublein et al., 2022, S. 2) haben. Nicht berücksichtigt werden in dieser Definition ein erfolgloses Zweitstudium (bei erfolgreich abgeschlossenem Erststudium) oder

ein weiteres Masterstudium sowie ein Wechsel des Studienfachs oder der Hochschule (Heublein et al., 2022). Diese Definition kann auch für Masterstudierende adaptiert werden. Demnach werden als Studienabbrecher (Master) Studierende verstanden, die ihr Masterstudium nicht erfolgreich abschließen, unberücksichtigt bleiben auch hierbei Studienfach- oder Hochschulwechsel (Heublein et al., 2022). Im Rahmen dieser Arbeit liegt der Fokus auf Bachelorstudierenden, dementsprechend erfolgt keine Unterscheidung zwischen Bachelor- und Masterstudierenden, die Begriffe Studienabbrecher und Studienabbruchquote beziehen sich somit auf Bachelorstudiengänge.

Studien des *Deutsches Zentrum für Hochschul- und Wissenschaftsforschung* (DZHW) zeigen seit vielen Jahren, dass rund ein Drittel aller Studienanfänger ihr Bachelorstudium vorzeitig ohne einen entsprechenden Abschluss abbricht (Heublein & Schmelzer, 2018; Heublein et al., 2014, 2017, 2020, 2022). Demnach brechen 35% der Studierenden ihr Studium vorzeitig ab – alle Studienabbruchquoten beziehen sich auf die Absolventen 2020 (Bachelor an Universitäten) in Relation zu den Studienanfängern 2016/2017. Die Fachrichtungen ›Mathematik/Naturwissenschaften‹ (50%), ›Geisteswissenschaften/Sport‹ (49%) und ›Ingenieurwissenschaften‹ (35%) weisen hier die höchsten Werte auf (Heublein et al., 2022). Als Ingenieurwissenschaften fasst das DZHW die Studiengänge Architektur, Bauingenieurwesen, Elektrotechnik, Informatik und Maschinenbau zusammen (Heublein et al., 2022). Der Studiengang Bauingenieurwesen weist mit 46% den höchsten Wert der Ingenieurwissenschaften auf (Heublein et al., 2022). Zur Verdeutlichung: Jeder zweite Studienanfänger im Bauingenieurwesen beendet sein Studium ohne erfolgreichen Abschluss, ein verheerender Zustand insbesondere im Hinblick auf die sinkenden Studienanfängerzahlen und den Fachkräftemangel in den entsprechenden Bereichen (siehe Kap. 2.2).

Bisher liegt kein umfassendes und allgemeingültiges Modell des Studienabbruchs vor (Nolden, 2019). Erste Modelle des Studienabbruchs wurden von Spady (1970) und Tinto (1975) vorgestellt, sie dienen bis heute als Grundlage für neuere Modelle. Konkret beschreibt Tinto (1975) in seinem auf Spady (1970) aufbauenden Längsschnittmodell, dass die Entscheidung zum Studienabbruch im Laufe der Zeit entsteht und nicht

zu einem bestimmten Zeitpunkt getroffen wird. Individuelle Merkmale (familiärer Hintergrund, Fähigkeiten/Fertigkeiten, Schulbildung), Ziele und Verpflichtungen sowie institutionelle Erfahrungen, die sich wiederum in akademischer bzw. sozialer Integration niederschlagen, beeinflussen diese Entscheidung (Tinto, 1975). Diese Kategorien werden auch in neueren Modellen weitgehend berücksichtigt (u.a. Bean & Metzner, 1985; Bornkessel, 2018; Gold, 1988; Heublein et al., 2017; Nolden, 2019; Ströhlein, 1983). Für diese Arbeit findet das Studienabbruchmodell des DZHW Anwendung (Heublein et al., 2017).

Dieses besagt, dass der Entscheidung zum Studienabbruch ein »mehrdimensionaler Prozess zugrunde [liegt], der in verschiedenen Phasen durch unterschiedliche Faktoren beeinflusst wird« (Heublein et al., 2017, S. 11). Dieser mehrdimensionale Prozess ist in Abb. 2.1 dargestellt. Die Abbildung zeigt, dass die drei Phasen (Studienvorphase, aktuelle Studiensituation und Entscheidung) von unterschiedlichen Faktoren beeinflusst werden. Die wichtigsten Faktoren, die zum Studienabbruch in ingenieurwissenschaftlichen Studiengängen führen, sind Leistungsprobleme (38%), mangelnde Studienmotivation (17%) und eine praktische Tätigkeit (18%) (Heublein et al., 2017). Andere Faktoren wie persönliche Gründe (9%), die finanzielle Situation (7%), die Studienbedingungen (6%), die familiäre Situation (3%), berufliche Alternativen (3%) oder die Studienorganisation (0%) spielen eine eher untergeordnete Rolle (Heublein et al., 2017).

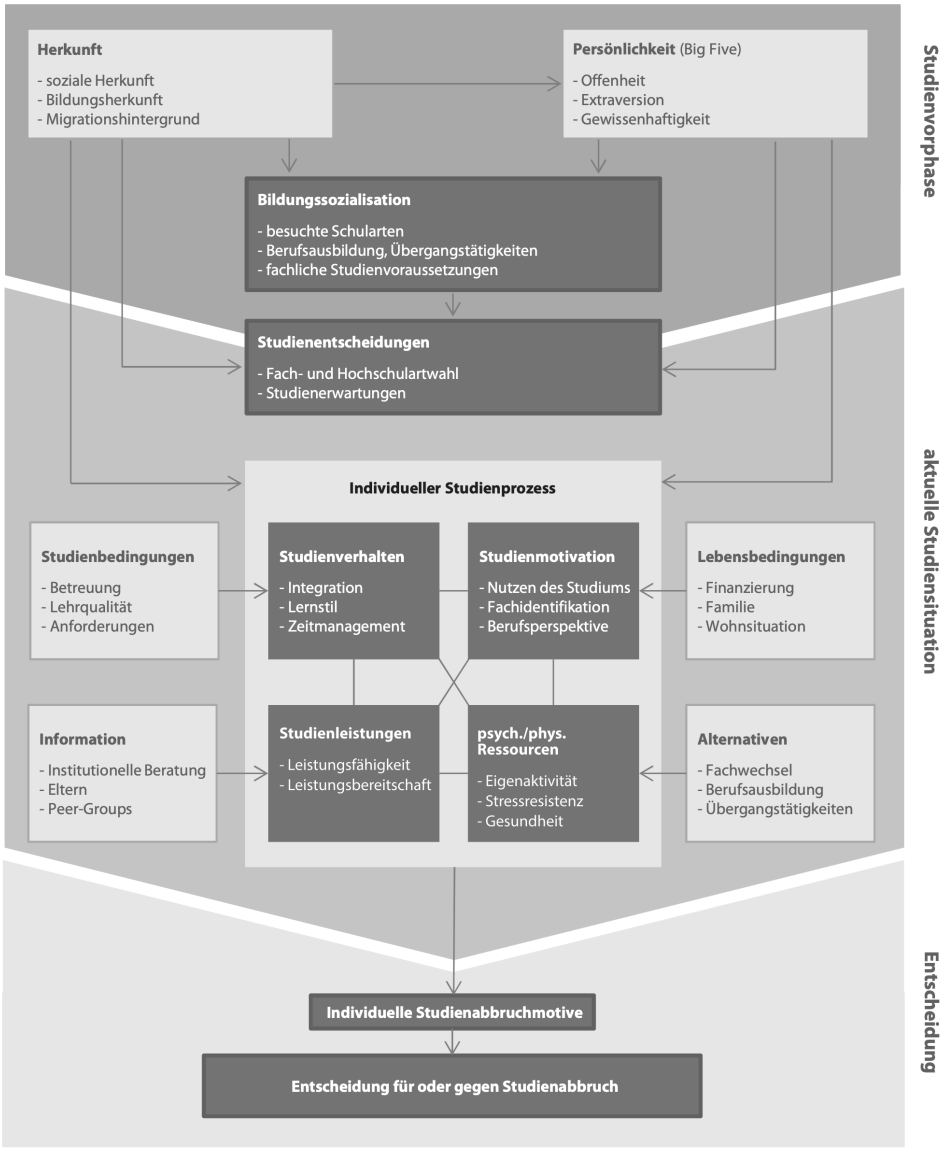
Die Ursachen der Leistungsprobleme lassen sich u.a. differenzieren in endgültig nicht bestandene Prüfungen (11%), eine Wahrnehmung von zu hohen Studienanforderungen (8%) oder auch Zweifel an der persönlichen Eignung des Studienfachs (8%) (Heublein et al., 2017).

Eine falsche Erwartungshaltung an das aufgenommene Studium (10%) erweist sich als primärer Auslöser für eine mangelnde Studienmotivation (Heublein et al., 2017). Diese falsche Erwartungshaltung lässt sich charakterisieren durch aus Studierendensicht »falsche[n] Vorstellungen von Fachinhalten und Anforderungen der selbstständigen Studienorganisation als auch auf Fehleinschätzungen ihres eigenen Leistungsvermögens« (Heublein et al., 2017, S. 29).

Als Hauptgrund für den Oberbegriff ›praktische Tätigkeit‹ geben

10% der Studienabbrecher den Wunsch nach einer praktischen Tätigkeit als ausschlaggebend für die vorzeitige Beendigung des Studiums an (Heublein et al., 2017).

Abbildung 2.1
Modell des Studienabbruchprozesses (Heublein et al., 2017, S. 12)



Die genannten Faktoren führen in der Regel in den ersten vier Fachsemestern zum Studienabbruch. Erste Gedanken an einen Studienabbruch haben 80% der Studienabbrecher bereits in den ersten drei

Fachsemestern, während die Abbruchentscheidung bei 50% bereits im zweiten Fachsemester getroffen wird (Heublein et al., 2017). Dies führt dazu, dass 42% der Studienabbrecher in den Ingenieurwissenschaften ihr Studium bereits in den ersten beiden Fachsemestern beenden, bei weiteren 31% erfolgt der Studienabbruch im 3. oder 4. Fachsemester (Heublein et al., 2017). Damit haben 73% der Studienabbrecher ihr ingenieurwissenschaftliches Studium in den ersten vier Fachsemestern ohne erfolgreichen Abschluss beendet, insgesamt beträgt die durchschnittliche Studiendauer bis zum Studienabbruch 4.3 Fachsemester (Heublein et al., 2017).

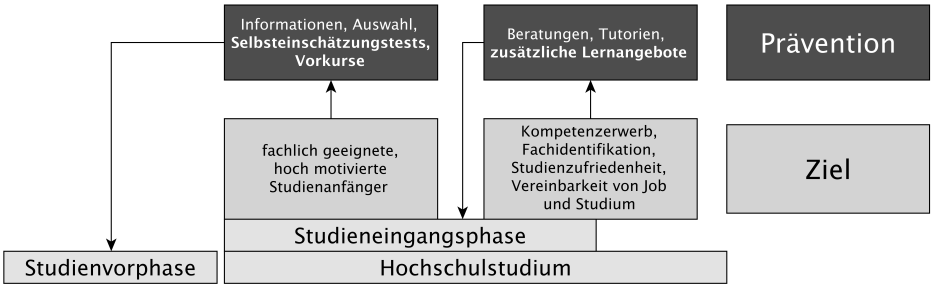
Auch nach Henn und Polaczek (2007) ist das vorzeitige Beenden des Studiums von Studierenden, die den größten Anteil der Exmatrikulationen beschreiben, auf mangelnden Studienerfolg in den ersten Semestern zurückzuführen. In ingenieurwissenschaftlichen Studiengängen kann die Studieneingangsphase somit zusammenfassend als entscheidend für den weiteren Erfolg der Studierenden im Verlauf ihres Studiums betrachtet werden.

Es stellt sich die Frage, worauf die Leistungsprobleme – die häufig in den Grundlagenfächern (Technische Mechanik oder Mathematik) auftreten (Heublein et al., 2010) – zurückzuführen sind. Eine mögliche Ursache für die beschriebenen Leistungsprobleme ist ein Rückgang spezieller fachlicher, auch mathematischer Kenntnisse bei Studienanfängern (u.a. Henn & Polaczek, 2007; Heublein & In der Smitten, 2013; Lehmann, 2018; Martschink, 2013; J. Müller et al., 2018). Dies erweist sich als besonders problematisch, da Ballard und Johnson (2004) und Hell et al. (2008) Zusammenhänge zwischen den grundlegenden mathematischen Kenntnissen und dem Erfolg in Mathematikkursen nachweisen konnten. Weitere Studien belegen, dass mathematische Kompetenzen als Prädiktor für die Studienabbruchsintenion dienen können (Fleischer et al., 2019; J. Müller et al., 2018). Kritisch ist zudem, dass die Wissenslücken in der Studieneingangsphase nur mit großen Schwierigkeiten geschlossen werden können (u.a. Dammann & Lang, 2019; Derr et al., 2017; Martschink, 2013; Willige et al., 2014).

Dieser Problematik haben sich Heublein und In der Smitten (2013) angenommen und das ›Referenzmodell zur Qualitätssicherung an Fakul-

täten der Ingenieurwissenschaften‹ entwickelt. Dieses Referenzmodell ist primär für die Bachelorphase der Studiengänge Maschinenbau und Elektrotechnik konzipiert, aufgrund der inhaltlichen Nähe aber auch auf weitere ingenieurwissenschaftliche Studiengänge wie das Bauingenieurwesen übertragbar. Heublein und In der Smitten (2013) weisen ausdrücklich darauf hin, dass es sich nicht um ein allgemeingültiges Modell des Qualitätsmanagements in Studium und Lehre handelt, da eine einheitliche Übertragbarkeit angesichts unterschiedlicher Voraussetzungen an verschiedenen Standorten nicht möglich, aber auch nicht notwendig ist. Das Referenzmodell soll die Hochschulen dabei unterstützen, die »Studienqualität so zu sichern und bei Bedarf zu erhöhen, dass die Studierenden in jeder Phase ihres Studiums wirkungsvoll begleitet werden, ihre Studienmotivation aufrecht erhalten und gefördert wird und ihr Fachwissen und ihre Kompetenzen erweitert werden« (Heublein und In der Smitten, 2013, S. 7). Generell lässt sich aus dem Referenzmodell ableiten, dass ein Bündel von Unterstützungsmaßnahmen zu unterschiedlichen Zeitpunkten im Studienverlauf zur Verbesserung des Studienerfolges hilfreich sein kann (Heublein & In der Smitten, 2013). Diese präventiv einzusetzenden Maßnahmen setzen in der Studienvor- und Studieneingangsphase an: In der Studienvorphase (vor Beginn des Studiums) können Selbsteinschätzungstests (engl.: *Online-Self-Assessments* [OSA]) und so genannte Vor- und Brückenkurse bzw. Online-Vorkurse (OV) hilfreich sein, in der Studieneingangsphase zusätzliche Lehrangebote wie z.B. iOM (Abb. 2.2).

Abbildung 2.2
Referenzmodell – Qualitätssicherung im Studienverlauf nach Heublein und In der Smitten (2013) (Pelz et al., 2019, S. 99)



Orientierungs- und Unterstützungsmaßnahmen, die Studienanfängern den Einstieg in das Studium erleichtern sollen, werden an vielen Universitäten bereits in einer großen Vielfalt und Zahl angeboten (Key & Hill, 2018). Es zeigt sich jedoch, dass Studieninteressierte bzw. Studienanfänger diese Angebote nur selten in Anspruch nehmen und dass insbesondere diejenigen, die eine solche Unterstützung dringend bräuchten, diesen Angeboten fern bleiben (u.a. Heublein et al., 2017; Tieben, 2019). Als Grund für dieses Phänomen nennen Voßkamp und Laging (2014) eine Überschätzung der eigenen Mathematikkenntnisse.

Die Bewertung solcher Unterstützungsmaßnahmen in der Studienvor- bzw. Studieneingangsphase ist hingegen uneindeutig. Bislang konnte empirisch »kein Zusammenhang zwischen der Teilnahme an einzelnen Angeboten zu Studienbeginn und positiven Erfahrungen in den ersten Wochen des Studiums« (Heublein et al., 2017, S. IX) nachgewiesen werden. Dennoch warnen Heublein et al. (2017) im selben Bericht ausdrücklich davor, daraus zu schließen, dass solche Angebote nicht wirksam und überflüssig seien. Es wäre mit einer deutlich höheren Zahl von Studienabbrüchen zu rechnen, wenn es solche Angebote nicht gäbe (Heublein et al., 2017).

2.3 Digitalisierung in der Hochschullehre

Unterstützungsangebote an Hochschulen können vielfältig sein. Im Folgenden werden drei Varianten vorgestellt, die auch in FUNDAMENT Anwendung finden. Alle drei Angebote wurden in FUNDAMENT als digitale Angebote umgesetzt, weshalb sich der Forschungsstand auch auf die digitalen Umsetzungen bezieht. Bei den digitalen Elementen handelt es sich um OSA, OV und iOM.

2.3.1 Online-Self-Assessment

Unter Selbsteinschätzungstests (OSA) sind online Studierfähigkeitstests zu verstehen, mit deren Hilfe Studierende vor Studienbeginn ihre Eignung (Leistungsstand) für einen bestimmten Studiengang an einer Hochschule prüfen können (Schmidt-Atzert et al., 2012). Die Ergebnisse

werden nur den potenziellen Studienanfängern mitgeteilt, die Hochschulen erhalten keine individuellen Ergebnisse und können diese somit nicht zur Selektion der Studieninteressierten nutzen (Schmidt-Atzert et al., 2012). Höft et al. (2020) charakterisieren OSA ausführlicher durch sieben Definitionsmerkmale:

1. Zielgruppe ›Studieninteressierte‹
2. internetbasiert (zeit- und ortsunabhängig)
3. frei zugänglich und in der Regel kostenlos
4. die Studieninteressierten erhalten realistische Informationen über die Tätigkeiten mit entsprechenden Aufgabenstellungen, die Art und Schwierigkeitsgrad der Studieninhalte und -bedingungen widerspiegeln
5. studiumsrelevante Personeneigenschaften können mittels psychodiagnostischer Verfahren erfasst werden
6. direkte Rückmeldung der Ergebnisse bezogen auf die Studienanforderungen
7. Bearbeitung ist freiwillig, die Ergebnisse dürfen von Hochschulen nicht für Zulassungsverfahren verwendet werden (Höft et al., 2020)

OSA lassen sich in zwei Gruppen unterteilen: Während sich allgemeine OSA an unentschlossene Studieninteressierte richten, sind fachspezifische OSA für Studieninteressierte konzipiert, die sich bereits für ein bestimmtes Studium entschieden haben (Höft & Fischer, 2023). Empirische Befunde bzgl. der Wirksamkeit von OSA liegen bisher noch nicht vor. Einzig erste Beschreibungen von Studierenden, die sich bereits im Studium befinden, auf Datenbasis eines eingesetzten OSA liegen in der Literatur vor (Schmitt et al., 2018).

Ambivalent sind die Ergebnisse einer Befragung von Studierenden des DZHW (Willige et al., 2014). Während rund 45% der befragten Studierenden der Ingenieurwissenschaften derartige Eingangs- und Selbsteinschätzungstests als ›nützlich‹ oder gar ›sehr nützlich‹ bezeichnen,

haben gerade einmal 25% der Befragten dieses Angebot genutzt (Willige et al., 2014).

Mittlerweile existieren eine Vielzahl unterschiedlicher (fachspezifischer) OSA, eine gute Übersicht für den deutschsprachigen Raum liefert das DACH-Gemeinschaftsprojekt OSA-Portal (o. J.) (Höft et al., 2020). In der Vergangenheit wurden zumeist OSA entwickelt, die fachunspezifische Themenfelder überprüfen, so werden bspw. für ein ingenieurwissenschaftliches Studium relevante Themen wie naturwissenschaftliche Zusammenhänge oder ingenieurwissenschaftliche Kontexte – u.a. derer der Technischen Mechanik – vernachlässigt, dagegen wird hauptsächlich der Themenkomplex der Mathematik in den Fokus gerückt (u.a. Bundesagentur für Arbeit [BA], o. J.; Duale Hochschule Baden-Württemberg [DHBW], o. J.; HafenCity Universität Hamburg [HCU Hamburg], o. J.; Technische Universität Hamburg [TUHH], o. J.; Universität Kassel, o. J.). Eine Rückmeldung über studienrelevante ingenieurwissenschaftliche Vorkenntnisse kann daher nur eingeschränkt gegeben werden. Diese Situation hat sich inzwischen geändert, da verschiedene OSA existieren, die in unterschiedlichem Maße auch ingenieurwissenschaftliche Kontexte (technisches Grundverständnis oder physikalische Grundlagen) einbeziehen (u.a. Ernst-Abbe-Hochschule Jena [EAH Jena], o. J.; Hochschule für Angewandte Wissenschaften Hamburg [HAW Hamburg], o. J.; Hochschule für Technik und Wirtschaft Berlin [htw Berlin], o. J.; Hochschule für Technik und Wirtschaft des Saarlandes [htw saar], o. J.; Hochschule Niederrhein [HS Niederrhein], o. J.; Rheinisch-Westfälischen Technischen Hochschule Aachen [RWTH Aachen], o. J.; Technische Hochschule Würzburg-Schweinfurt [THWS], o. J.).

Die Evaluationsergebnisse zu den allgemeinen OSA fassen Thiele und Kauffeld (2019) zusammen: Studienanfänger akzeptieren die OSA, nutzen sie intensiv, bewerten sie überwiegend sehr positiv, erhalten einen zusätzlichen Informationsgewinn sowie Erwartungsklarheit und Sicherheit, den richtigen Studiengang gewählt zu haben. Darüber sind die erzielten Ergebnisse in der Lage, die späteren Studienleistungen und die Studienzufriedenheit vorherzusagen (Thiele & Kauffeld, 2019). Allerdings beziehen sich diese Ergebnisse auf allgemeine OSA, für fachspezifische OSA liegen keine eindeutigen Forschungsergebnisse vor.

2.3.2 Online-Vorkurs

Meist in Kombination mit den im vorigen Kapitel (2.3.1) genannten fachspezifischen OSA wird mit den sogenannten Online-Vorkursen (OV) ein weiteres Unterstützungsangebot in der Studienvorphase eingesetzt. Vorkurse, auch Brückenkurse genannt, in Präsenzform zur Auffrischung oder Vertiefung von Schulwissen (Tieben, 2019) haben an Hochschulen eine lange Tradition. Abel und Weber (2014) berichtet von einem ersten mathematischen Vorkurs, der an der Hochschule Esslingen im WiSe 1983/1984 als ›Therapie‹ einem Multiple-Choice-Wissenstest – im weitesten Sinne vergleichbar mit einem OSA – nachgeschaltet war. Klassische Vorkurse sind als Präsenzveranstaltungen in der Studienvorphase angesiedelt, finden über einen Zeitraum von ca. zwei Wochen statt und orientieren sich thematisch an den Inhalten der Oberstufenmathematik, die Teilnahme ist in der Regel freiwillig (Kempen & Wassong, 2017; Tieben, 2019).

Vorkurse werden an 75% der Hochschulen angeboten, an Universitäten liegt der Wert mit 72% etwas niedriger, während in den Ingenieurwissenschaften Vorkurse an 87% der Hochschulen angeboten werden (Falk & Marschall, 2021). Knapp 63% der Studierenden in den Ingenieurwissenschaften (Universitäten) halten Vorkurse zur Aufarbeitung fachlicher Wissenslücken und Voraussetzungen für ›nützlich‹ oder sogar ›sehr nützlich‹ (Willige et al., 2014). Dieser Wert deckt sich weitgehend mit den rund 52% der Studierenden, die angeben, einen Vorkurs besucht zu haben, und liegt deutlich über den 37% der Studierenden in den naturwissenschaftlichen Studiengängen, die den zweitgrößten Anteil ausmachen (Willige et al., 2014). Falk und Marschall (2021) weisen für die Ingenieurwissenschaften Zahlen in ähnlicher Größenordnung aus (58%). Diese Fakten verdeutlichen den hohen Stellenwert von Vorkursen in den Ingenieurwissenschaften.

Um die Teilnehmerzahlen weiter zu erhöhen, haben einige Hochschulen die Studienvorphase mit unterschiedlichen Ansätzen neu konzipiert, sodass u.a. klassische Präsenzvorkurse in E-Learning-Angebote (OV) überführt wurden (Beispiele für mathematische Vorkurse und OV geben Bausch et al., 2014). Aus mehreren Verbundprojekten sind so OV entstanden, die von den Hochschulen empfohlen werden, als Beispiele

seien hier das *Virtuelle[s] Eingangstutorium Mathematik, Informatik, Naturwissenschaften, Technik* (VEMINT) (Universität Paderborn, o. J.) oder der *Online Mathematik Brückenkurs* (OMB+) (integral-learning GmbH, o. J. b) genannt (weitere Mathematik-OV: integral-learning GmbH, o. J. a; MINT-Kolleg Baden-Württemberg, o. J. a). Der Schwerpunkt der bestehenden OV liegt jedoch auf der Vermittlung mathematischer Grundlagen für Studieninteressierte bzw. Studienanfänger. Ingenieurwissenschaftliche Kontexte werden selten aufgegriffen, lediglich physikalische Grundlagen sind in wenigen OV vorhanden (Hochschule RheinMain, o. J.; MINT-Kolleg Baden-Württemberg, o. J. b; Technische Universität Dresden [TU Dresden], o. J.).

Evaluationsergebnisse zur Wirksamkeit und insbesondere zu Langzeiteffekten (u.a. in Bezug auf Studienabbruch) von Vorkursen leiden zumeist an methodischen Mängeln, liefern widersprüchliche Ergebnisse und lassen keine generalisierbaren Aussagen zu (Tieben, 2019). Tieben (2019) untersucht Vorkurse in ingenieurwissenschaftlichen Studiengängen. Eine differenzierte Betrachtung nach Fachhochschulen und Universitäten zeigt, dass die Teilnahme an einem Vorkurs an Fachhochschulen keinen Einfluss auf den Studienabbruch hat (Tieben, 2019). An Universitäten hingegen führt die Teilnahme zu einem geringeren Studienabbruch (Tieben, 2019). Allerdings schränkt die Autorin die Aussagekraft der Ergebnisse ein, weil sie leicht überschätzt werden, da die Teilnehmenden an Vorkursen in der Regel etwas motivierter und leistungsstärker sind und somit ein Studienabbruch generell unwahrscheinlicher ist (Tieben, 2019). Zudem stellt Tieben (2019) die Frage, ob die soziale Integration bzw. die Studienbindung durch die Teilnahme an einem Vorkurs erhöht wird und damit einen Einfluss auf den Studienabbruch hat. Entsprechende Studien zu OV sind bislang rar. In einer der wenigen Studien evaluiert Fischer (Fischer, 2014) die Wirksamkeit des – im Blended-Learning-Konzept durchgeführten – OV VEMINT. Er kann nachweisen, dass die Teilnehmenden des Blended-Learning-VEMINT-Vorkurses zwar tendenziell bessere Leistungen erzielen als die Teilnehmenden des mathematischen Präsenzvorkurses, insgesamt aber nur ein geringer Leistungszuwachs erkennbar ist (Fischer, 2014). Auswirkungen auf einen möglichen Studienabbruch prüft Fischer (2014)

nicht.

2.3.3 interaktive Online-Module

Während OSA und OV in der Studienvorphase angesiedelt sind, ist auch die Studieneingangsphase entscheidend für den Studienerfolg. Insbesondere in der Studieneingangsphase haben Studienanfänger Schwierigkeiten, die zentralen Konzepte des Faches zu verstehen (B. G. Prusty & Russell, 2011). Diese Verständnisprobleme können zu den bereits beschriebenen Leistungsproblemen führen. Die Verknüpfung von Vorlesungsinhalten mit den entsprechenden Kernkonzepten stellt somit ein Problem für die Studierenden dar. Der Einsatz von z.B. Demonstrationsexperimenten innerhalb der Veranstaltung als zusätzliche Visualisierung der theoretischen Sachverhalte führt nicht zu einer Verbesserung (Crouch et al., 2004). Derartige Untersuchungen gibt es allerdings bisher nur für die Physik und nicht im ingenieurwissenschaftlichen Kontext.

Die Förderung des konzeptionellen Wissens kann durch den individuellen Dozenten-Studierenden-Diskurs gefördert werden, wofür an vielen Hochschulen verschiedene Angebote wie Hörsaalübungen oder Tutorien existieren, deren Teilnehmeranzahl ist jedoch stark begrenzt ist. Um dieser Problematik entgegenzuwirken, wurden vor allem im angelsächsischen Raum zusätzliche Lernangebote, sogenannte interaktive Online-Module (iOM), entwickelt (B. G. Prusty et al., 2009; B. G. Prusty & Russell, 2011; G. Prusty et al., 2011). Verallgemeinerbare Ergebnisse liegen jedoch noch nicht vor. Dies gilt auch für den Themenkomplex der Technischen Mechanik.

2.4 Fachspezifische Inhalte

Wie bereits in Kap. 2.2.1 gezeigt werden konnte, stellen die fachspezifischen Inhalte der Technischen Mechanik und der Mathematik insbesondere Studienanfänger vor große Schwierigkeiten. In den folgenden beiden Kapiteln soll daher erläutert werden, welche Inhalte der Technischen Mechanik in den entsprechenden Lehrveranstaltungen der ersten beiden Fachsemester vermittelt werden. Diese Inhalte müssen auch in den fachspezifischen Tests entsprechend berücksichtigt werden,

damit die Testinstrumente den Studienerfolg (Klausurnoten) vorhersagen können. Bei der Mathematik wird sich auf die Rechenfähigkeit beschränkt, d.h. auf mathematische Grundkenntnisse, die zu Beginn eines Ingenieurstudiums vorausgesetzt werden.

2.4.1 Technische Mechanik

Die Mechanik ist ein Teilgebiet der Physik, welches sich mit der Wirkung von Kräften auf Körper befasst (Assmann & Selke, 2010; Gross et al., 2019). Sie ist das älteste und damit auch das am weitesten entwickelte Teilgebiet der Physik (Gross et al., 2019). Nach Gross et al. (2019) gewinnt die Mechanik als Grundlage der Technik weiter an Bedeutung, da immer neue Anwendungsgebiete erschlossen und komplexe Problemstellungen exakt analysiert werden können. Grundlage der Mechanik sind Naturgesetze mit axiomatischem Charakter, d.h. »Aussagen, die vielfachen Beobachtungen entnommen sind und aus der Erfahrung heraus als richtig angesehen werden; auch ihre Folgerungen werden durch die Erfahrung bestätigt« (Gross et al., 2019, S. XIII). Diese Naturgesetze und auch ihre Anwendung werden als Modelle ergründet, die mechanische Charakteristika des realen Körpers oder Systems aufweisen (Gross et al., 2019). Die Modelle (Systeme und Zusammenhänge der Realität) werden mit mathematischen Formeln beschrieben, sodass neben den physikalischen auch mathematische Grundlagen für das Verständnis der Mechanik wichtig sind und den Studienanfängern den Zugang zur Mechanik zusätzlich erschweren (Magnus & Müller-Slany, 2009).

In der Mechanik wird zwischen der Analytischen Mechanik und der Technischen Mechanik unterschieden (Gross et al., 2019; Magnus & Müller-Slany, 2009). Die Analytische Mechanik hat zum Ziel, grundlegende Erkenntnisse und Gesetzmäßigkeiten aus mechanischen Prozessen abzuleiten (Gross et al., 2019). Werden dagegen die physikalischen Erkenntnisse der Mechanik auf technische Objekte oder Prozesse angewendet, wird von Technischer Mechanik gesprochen (Assmann & Selke, 2010), also wenn »die Probleme und Ansprüche der konstruierenden und berechnenden Ingenieure« (Gross et al., 2019, S. XIV) im Vordergrund stehen.

Die Inhalte der Mechanik lassen sich nach verschiedenen Gesichts-

punkten gliedern. Magnus und Müller-Slany (2009) nennen ein Ordnungssystem, das auf den Eigenschaften der betrachteten Körper beruht, wobei die Grenzen nicht klar trennbar sind:

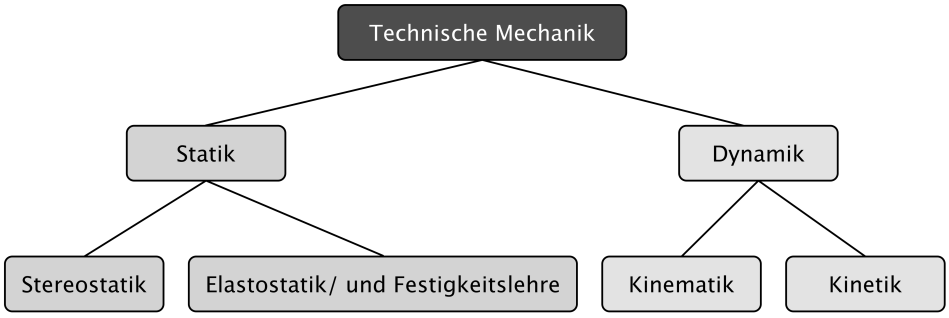
- Stereo-Mechanik (Punktmassen und starre Körper)
- Elasto-Mechanik (elastische Körper)
- Plasto-Mechanik (plastische Körper)
- Fluid-Mechanik (flüssige und gasförmige Körper) (Magnus & Müller-Slany, 2009, S. 15).

Eine weitere Möglichkeit der Systematisierung ist die Gliederung nach Aggregatzuständen des Körpers, hier erfolgt die Einteilung nach der Mechanik fester, flüssiger (Hydromechanik) oder gasförmiger (Aero- und Gasdynamik) Körper (Gross et al., 2019; Schröder, 2020). Die Mechanik fester Körper wird wiederum nach der Grundfunktion der Mechanik (Untersuchung von Kräften und Bewegungen) in Kinematik und Dynamik unterteilt (Gross et al., 2019). Die Kinematik ist die Lehre von der Bewegung fester Körper ohne Berücksichtigung der Kräfte, die ihrerseits die Bewegung hervorrufen oder beeinflussen können (Gross et al., 2019; Schröder, 2020). Im Gegensatz dazu beschäftigt sich die Dynamik mit dem Gleichgewicht (Schröder, 2020), bei dem die Kräfte berücksichtigt werden und in Beziehung zur Bewegung des festen Körpers stehen (Gross et al., 2019). Die Dynamik kann wiederum in Kinetik und Statik unterteilt werden (Gross et al., 2019; Magnus & Müller-Slany, 2009; Schröder, 2020). Während die Kinetik die Beziehungen zwischen Kräften und Bewegungen beschreibt (Gross et al., 2019; Schröder, 2020), beschäftigt sich die Statik mit dem Gleichgewicht ruhender Körper (Assmann & Selke, 2010; Schröder, 2020). Zur Statik gehören die Stereostatik (Statik starrer Körper), die Elastostatik/Festigkeitslehre (Statik deformierbarer Körper) sowie die Hydrostatik (Statik flüssiger Körper) (Schröder, 2020).

Damit ist die Gliederung der Mechanik erfolgt. Dammann (2016) führt dagegen in seiner Dissertation an, dass die genannte Gliederung nur als Basis für die gelehrte Technische Mechanik zu verstehen ist. In der Lehrpraxis an deutschen Hochschulen wird die Technische Mechanik

wie in Abb. 2.3 dargestellt unterteilt (Dammann, 2016). Während die Gliederung der Mechanik zunächst die bewegten und erst anschließend die ruhenden Körper beschreibt, wird in der Praxis die Technische Mechanik in umgekehrter Reihenfolge behandelt (Dammann, 2016). In den ersten Fachsemestern stehen die ruhenden Körper mit der Statik und Elastostatik/Festigkeitslehre im Mittelpunkt und erst in den darauffolgenden Semestern wird die Dynamik behandelt (Dammann, 2016).

Abbildung 2.3
Gliederung der Technischen Mechanik (in Anlehnung an Dammann, 2016, S. 69)



U.a. in der Lehrbuchreihe ›Technische Mechanik‹ (Gross et al., 2019, 2021a, 2021b, 2023) findet sich dieser Ansatz wieder. Als Mitautor der ersten drei Bände (Gross et al., 2019, 2021a, 2021b) orientiert sich auch Schröder (2020, 2021) an dieser Einteilung, die er auch in den Skripten zu seinen Vorlesungen für Studienanfänger des Bauingenieurwesen an der UDE in den ersten beiden Fachsemestern anbietet:

- Mechanik 1 - Stereostatik (1. Fachsemester)
  - Das zentrale Kräftesystem
  - Allgemeine (nichtzentrale) Kräftegruppe und Gleichgewicht des starren Körpers
  - Mittel- und Schwerpunktsberechnung
  - Lagerreaktion
  - Mehrteilige Tragwerke

  - Schnittgrößen am Balken, Rahmen und Bogen
  - Mechanische Arbeit und Potentialbegriff
  - Stabilität des Gleichgewichts (starrer Körper)
  - Coulombsche Theorie der Reibung (Schröder, 2020).

- Mechanik 2 - Elastostatik (2. Fachsemester)
  - Flächenträgheitsmomente
  - Statik deformierbarer Körper
  - Mehrdimensionale Spannungszustände
  - Verallgemeinertes Hookesches Gesetz
  - Technische Biegetheorie des dünnen prismatischen Balkens
  - Schubspannungen infolge Querkraft
  - Torsion
  - Arbeitsbegriff, Formänderungsenergie und Prinzip der virtuellen Kräfte
  - Stabilität des Gleichgewichts
  - Verbundquerschnitte
  - Biegetheorie gekrümmter Träger
  - Festigkeitshypothesen (Schröder, 2021).

Die genannten Themenbereiche müssen in den fachspezifischen Tests in der Studieneingangsphase abgebildet werden.

2.4.2 Mathematik

Als Hauptgrund für den Studienabbruch gelten Leistungsprobleme, die vor allem in den Grundlagenfächern auftreten (siehe Kap. 2.2.1). Eines der wichtigsten Grundlagenfächer in einem ingenieurwissenschaftlichen Studium und damit auch im Bauingenieurwesen, ist die Mathematik. Da in den Mathematikvorlesungen bereits neue, über das Schulwissen hinausgehende Inhalte vermittelt werden, müssen bestimmte grundlegende Rechenfähigkeit als bekannt vorausgesetzt werden.

Welche Themenbereiche der Mathematik als Mindeststandard vorausgesetzt werden können, wurde auf europäischer Ebene von der *European Society for Engineering Education* (SEFI) und auf nationaler Ebene von der Arbeitsgruppe *Cooperation Schule-Hochschule* (cosh) in Baden-Württemberg zusammengestellt.

In der mittlerweile dritten Auflage des ›A framework for mathematics curricula in engineering education: a report of mathematics working group‹ hat die SEFI das Curriculum der Mathematik (Ingenieurstudium) in vier Niveaustufen eingeteilt (Alpers, 2013). Diese Niveaus sollen die hierarchische Struktur der Mathematik während des Ingenieurstudiums darstellen, in der immer komplexere reale Anwendungen mit der Mathematik verknüpft werden (Alpers, 2013). Den Ausgangspunkt (1. Niveau) bildet der sogenannte *Core Zero* (Alpers, 2013). Dabei handelt es sich um realistische Vorkenntnisse, die vorausgesetzt werden dürfen, die aber aufgrund der unterschiedlichen Schulbildung nicht immer eingehalten werden können, weshalb es durchaus möglich ist, diese Themen in den ersten beiden Fachsemestern zu behandeln (Alpers, 2013). Die Inhalte von *Core Zero* sind so wesentlich, dass nur minimale Abweichungen zulässig sind (Alpers, 2013). In den weiteren Niveaustufen werden für die allgemeinen Ingenieurwissenschaften wesentliche Kenntnisse (*Level 1 - Kernbereiche [engl. core]*), spezielle und weiterführende Kenntnisse (*Level 2 - Wahlmodule [engl. electives]*) sowie hochspezialisierte Kenntnisse und Fertigkeiten (*Level 3 - Fachmodule [engl. specialist modules]*) berücksichtigt (Alpers, 2013). Da im Folgenden nur die mathematischen Mindestanforderungen für ein ingenieurwissenschaftliches Studium relevant sind, werden diese nach dem von der SEFI vorgeschlagenen *Core Zero* aufgelistet:

- Algebra
  - Arithmetik der reellen Zahlen
  - Algebraische Ausdrücke und Formeln
  - Lineare Gesetze
  - Quadratische und kubische Zahlen, Polynome

- Analysis
  - Funktionen und Umkehrfunktionen
  - Folgen, Reihen und binomische Formeln
  - Logarithmische- und Exponentialfunktionen
  - Extremwerte (Minima und Maxima)
  - Unbestimmte Integrale
  - Bestimmte Integrale und ihre Anwendung auf Flächen und Volumina
  - Komplexe Zahlen
  - Beweise
- Diskrete Mathematik
  - Mengen
- Geometrie und Trigonometrie
  - Koordinatengeometrie
  - Trigonometrische Funktionen und ihre Anwendung
  - Winkelfunktionen
- Statistik und Wahrscheinlichkeitsrechnung
  - Umgang mit (statistischen) Daten
  - Wahrscheinlichkeiten (Alpers, 2013, eigene Übersetzung).

Der »Mindestanforderungskatalog Mathematik« der Arbeitsgruppe cosh ist prinzipiell für den Übergang von einer baden-württembergischen Schule an eine Hochschule desselben Bundeslandes konzipiert – ist aber auch auf andere Bundesländer übertragbar – und gilt für das Studium der Studienfächer Wirtschaft, Mathematik, Informatik, Naturwissenschaft und Technik (WiMINT) (Cooperation Schule-Hochschule [cosh], 2021). Ein Vergleich mit dem *Core Zero* der SEFI (Alpers, 2013) zeigt, dass beide Anforderungskataloge viele Gemeinsamkeiten aufweisen. Ein Unterschied besteht darin, dass die Mindestanforderungen

nach der Arbeitsgruppe cosh (2021) im Wesentlichen auf dem in der Schule vermittelten Wissen aufbauen, während die Anforderungen der SEFI (Alpers, 2013) darüber hinausgehen (z.B. komplexe Zahlen) – diese Kenntnisse können aber, wie bereits erwähnt, im Laufe des ersten Studienjahres nachgeholt werden. Die im Kernlehrplan für die Sekundarstufe II für Gymnasien und Gesamtschulen in NRW genannten Themen – Inhaltsfelder Funktion und Analysis, Analytische Geometrie und Lineare Algebra sowie Stochastik (MSB NRW, 2023) – finden sich alle in den Katalogen der Mindestanforderungen wieder.

Wie bei den Inhalten der Technischen Mechanik müssen die genannten Mindestanforderungen auch bei den fachspezifischen Tests berücksichtigt werden. Durch Tests während der gesamten Studieneingangsphase kann überprüft werden, ob und wann die Studierenden die von der SEFI (Alpers, 2013) definierten Basisanforderungen des *Core Zero* des ersten Studienjahres (2. Fachsemester) erreichen.

2.5 Forschungsprojekte FUNDAMENT und ALSTER

In dieser Arbeit werden Daten aus zwei Forschungsprojekten zur Beantwortung der Forschungsfragen verwendet. Dabei handelt es sich um die Verbundforschungsprojekte FUNDAMENT und ein Teilprojekt der Forschergruppe ALSTER. Während die Datenerhebung von FUNDAMENT in Kap. 5 ausführlich dargestellt wird, beschränkt sich die Darstellung in Bezug auf ALSTER im wesentlichen auf die Verwendung der erhobenen Daten.

2.5.1 FUNDAMENT

Das Verbundforschungsprojekt *Förderung des individuellen Lernerfolgs mittels digitaler Medien im Bauingenieurstudium* (FUNDAMENT)[2] der UDE und der Technischen Universität Kaiserslautern hat zum Ziel

[2]FUNDAMENT wurde im Rahmen der Richtlinie zur Förderung von Forschung zur digitalen Hochschulbildung innerhalb der ersten Förderlinie im Forschungsfeld Digitale Hochschullehre vom BMBF gefördert (FKZ 16DHL1024).

individuelle Lernprozesse im Bauingenieurwesen durch den Einsatz digitaler Hochschullehre zu fördern. Durch die Entwicklung und den Einsatz geeigneter hochschuldidaktischer Maßnahmen soll die Qualität der universitären Lehr- und Unterstützungsangebote weiter verbessert und als Folge die Studienabbruchquote im Bauingenieurwesen gesenkt werden.

Das in FUNDAMENT entwickelte Förderkonzept beinhaltet – in Anlehnung an das Referenzmodell von Heublein und In der Smitten (2013) (siehe Kap. 2.2.1) – mehrere präventive Maßnahmen. Das Konzept kommt sowohl in der Studienvorphase, als auch in den ersten beiden Fachsemestern (Studieneingangsphase) in den Veranstaltungen der Technischen Mechanik zum Tragen. Im Einzelnen handelt es sich in der Studienvorphase um ein fachspezifisches OSA und ein OV, der die gleichen Themenbereiche aufgreift. Die Studieneingangsphase umfasst mit den sogenannten iOM ein ›Drei-Säulen-Konzept‹ in Form von Lernvideos, parametrisierten Online-Übungsaufgaben und der Möglichkeit zur Online-Kommunikation über Foren. Alle Online-Elemente sind in einer Moodle-Umgebung (Moodle Pty Ltd, o. J.) in Kombination mit dem serverbasierten System JACK (Goedicke, 2017) realisiert.

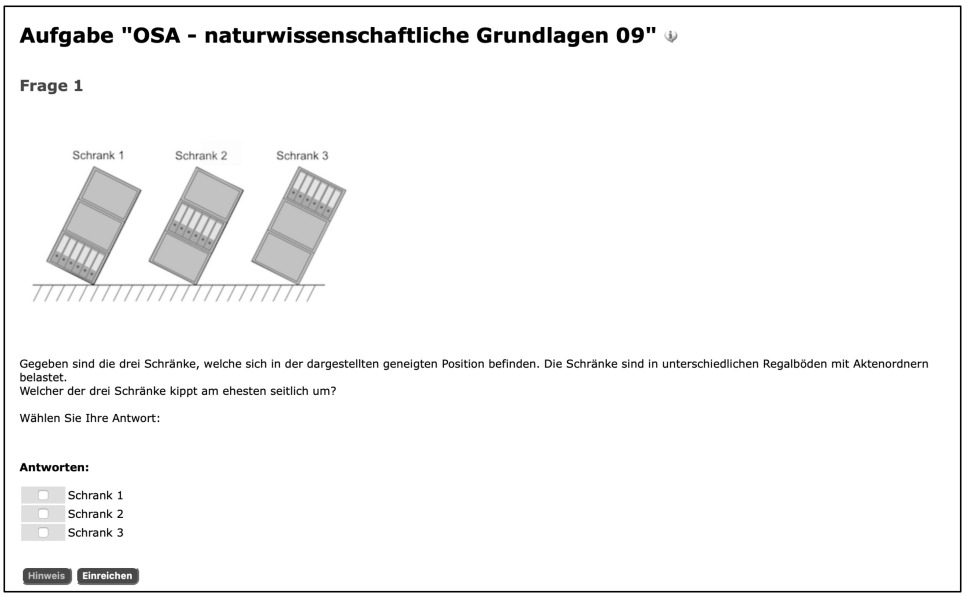
Im Rahmen dieser Arbeit werden ausschließlich die Ergebnisse der UDE vorgestellt. Weitere Details sowie bereits veröffentlichte Ergebnisse zu den entwickelten Online-Elementen finden sich in den beiden folgenden Kapiteln.

2.5.1.1 FUNDAMENT - Studienvorphase

Um den Studieninteressierten die Möglichkeit zu geben, ihre Vorkenntnisse mit den fachlichen Anforderungen des Studiums des Bauingenieurwesens abzugleichen, wird ein fachspezifisches OSA angeboten. Dieses OSA beinhaltet Vorwissenstests der Inhaltsbereiche der mathematischen (MG) und naturwissenschaftlichen Grundlagen (NG). Die in den ersten Studiensemestern relevanten Themen der Ingenieurmathematik (Kap. 2.4.2) werden in den mathematischen Grundlagen behandelt, hierzu zählen: Bruchrechnung, quadratische Funktionen, Gleichungssysteme, Trigonometrie, Vektorrechnung und Analysis. Dagegen werden in den naturwissenschaftlichen Grundlagen die Grundlagen der Physik vermit-

telt, die als Einführung in die Technische Mechanik betrachtet werden (Kap. 2.4.1), inhaltlich sind dies: Kräfte/Momente, Schwerpunkt, Lager, Stabilität und Zug, Druck, Biegung. Als Beispiele sind in Abb. 2.4 ein Screenshot eines Items zum Themengebiet ›Schwerpunkt‹ aus dem OSA naturwissenschaftliche Grundlagen und in Abb. 2.5 ein Screenshot eines Items aus dem Themengebiet ›Bruchrechnung‹ aus dem Bereich mathematische Grundlagen abgebildet.

Abbildung 2.4
Item des OSA naturwissenschaftliche Grundlagen zum Themengebiet ›Schwerpunkt‹



Nach Abschluss des OSA erhalten die Studieninteressierten ein individuelles Feedback, welches die Schwachstellen des Vorwissens benennt und entsprechend auf relevante Themengebiete im OV verweist (Abb. 2.6).

Der OV ist so konzipiert, dass die Studieninteressierten die im zuvor durchgeführten OSA aufgedeckten Wissenslücken aufarbeiten können. Um dies zu gewährleisten, ist der OV, wie bereits das OSA, thematisch in die Inhaltsbereiche naturwissenschaftliche und mathematische Grundlagen unterteilt. Die beiden Inhaltsbereiche sind didaktisch unterschiedlich aufbereitet. Während die mathematischen Grundlagen die Konzepte des ›Lernen am Beispiel‹ (Schworm, 2004) mit ›Informativem

Abbildung 2.5
Item des OSA mathematische Grundlagen zum Themengebiet ›Bruchrechnung‹

Aufgabe "OSA mathematische Grundlagen 02"

Frage 1

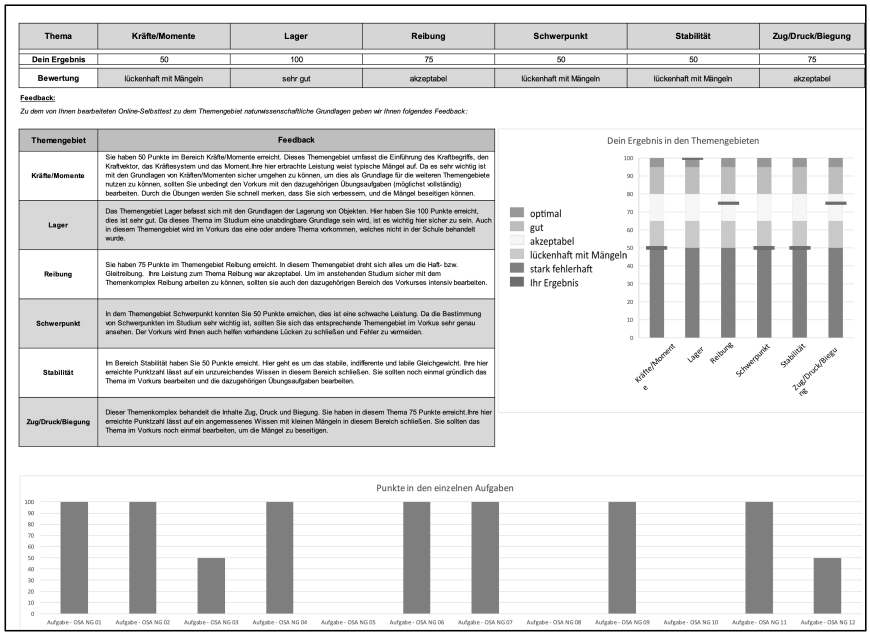
Gegeben ist folgender Term:

$$\frac{a}{b} = \left(2 \cdot \frac{2}{3} + \frac{8}{9}\right) \cdot \frac{4}{3}$$

Berechnen Sie die Werte für a und b. Vereinfachen Sie soweit wie möglich und tragen Sie Ihre Lösung ein.

a =
b =

Abbildung 2.6
Individuelles Feedback zum OSA naturwissenschaftliche Grundlagen (Pelz et al., 2021, S. 101)



Thema	Kräfte/Momente	Lager	Reibung	Schwerpunkt	Stabilität	Zug/Druck/Biegung
Dein Ergebnis	50	100	75	50	50	75
Bewertung	lückenhaft mit Mängeln	sehr gut	akzeptabel	lückenhaft mit Mängeln	lückenhaft mit Mängeln	akzeptabel

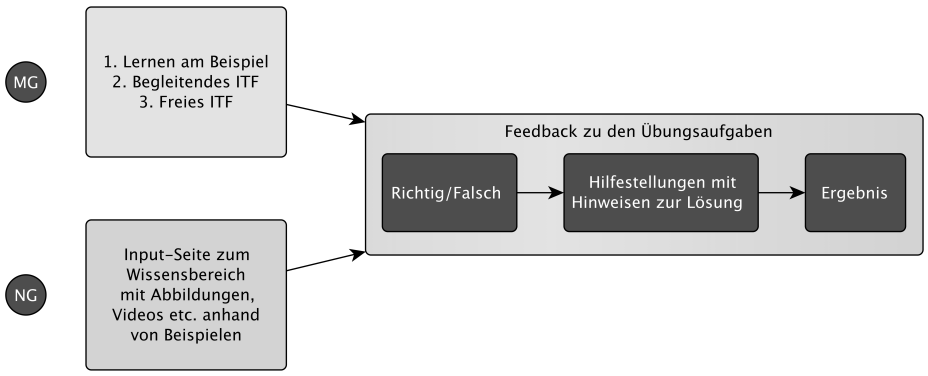
Feedback:

Zu dem von Ihnen bearbeiteten Online-Selbsttest zu dem Themengebiet naturwissenschaftliche Grundlagen geben wir Ihnen folgendes Feedback:

Themengebiet	Feedback
Kräfte/Momente	Sie haben 50 Punkte im Bereich Kräfte/Momente erreicht. Dieses Themengebiet umfasst die Einführung des Kraftbegriffs, den Kraftvektor, das Kräftesystem und das Moment.Ihre hier erbrachte Leistung weist typische Mängel auf. Da es sehr wichtig ist mit den Grundlagen von Kräften/Momenten sicher umgehen zu können, um dies als Grundlage für die weiteren Themengebiete nutzen zu können, sollten Sie unbedingt den Vorkurs mit den dazugehörigen Übungsaufgaben (möglichst vollständig) bearbeiten. Durch die Übungen werden Sie schnell merken, dass Sie sich verbessern, und die Mängel beseitigen können.
Lager	Das Themengebiet Lager befasst sich mit den Grundlagen der Lagerung von Objekten. Hier haben Sie 100 Punkte erreicht, dies ist sehr gut. Da dieses Thema im Studium eine unabdingbare Grundlage sein wird, ist es wichtig hier sicher zu sein. Auch in diesem Themengebiet wird im Vorkurs das eine oder andere Thema vorkommen, welches nicht in der Schule behandelt wurde.
Reibung	Sie haben 75 Punkte im Themengebiet Reibung erreicht. In diesem Themengebiet dreht sich alles um die Haft- bzw. Gleitreibung. Ihre Leistung zum Thema Reibung war akzeptabel. Um im anstehenden Studium sicher mit dem Themenkomplex Reibung arbeiten zu können, sollten sie auch den dazugehörigen Bereich des Vorkurses intensiv bearbeiten.
Schwerpunkt	In dem Themengebiet Schwerpunkt konnten Sie 50 Punkte erreichen, dies ist eine schwache Leistung. Da die Bestimmung von Schwerpunkten im Studium sehr wichtig ist, sollten Sie sich das entsprechende Themengebiet im Vorkus sehr genau ansehen. Der Vorkurs wird Ihnen auch helfen vorhandene Lücken zu schließen und Fehler zu vermeiden.
Stabilität	Im Bereich Stabilität haben Sie 50 Punkte erreicht. Hier geht es um das stabile, indifferente und labile Gleichgewicht. Ihre hier erreichte Punktzahl lässt auf ein unzureichendes Wissen in diesem Bereich schließen. Sie sollten noch einmal gründlich das Thema im Vorkurs bearbeiten und die dazugehörigen Übungsaufgaben bearbeiten.
Zug/Druck/Biegung	Dieser Themenkomplex behandelt die Inhalte Zug, Druck und Biegung. Sie haben in diesem Thema 75 Punkte erreicht.Ihre hier erreichte Punktzahl lässt auf ein angemessenes Wissen mit kleinen Mängeln in diesem Bereich schließen. Sie sollten das Thema im Vorkurs noch einmal bearbeiten, um die Mängel zu beseitigen.

Tutoriellen Feedback‹ (ITF) (Narciss, 2006) in einem ingenieurwissenschaftlichen Kontext berücksichtigen, basieren die naturwissenschaftlichen Grundlagen weitgehend auf einer klassischen Einstiegsseite zum Wissensgebiet (Input) mit Abbildungen und Beispielen sowie vertiefende Übungsaufgaben (Abb. 2.7).

Abbildung 2.7
FUNDAMENT – Aufbau des OV (in Anlehnung an Pelz et al., 2021, S. 102)



Anmerkungen. ITF = Informatives Tutorielles Feedback.

Exemplarisch ist in Abb. 2.8 ein Ausschnitt der Einführungsseite aus dem OV naturwissenschaftliche Grundlagen zum Themenbereich ›Kräfte‹ dargestellt.

Das gesamte Untersuchungsdesign der Studie FUNDAMENT ist klassisch nach dem Experimental-/Kontrollgruppendesign konzipiert, wobei aufgrund des Einsatzes der verschiedenen digitalen Elemente eine Unterteilung in Studienvor- und Studieneingangsphase möglich ist. Die Studienvorphase stellt den ersten MZP dar, hier soll primär die Wirksamkeit des OV überprüft werden. Die Kontrollgruppe durchläuft das OSA zweimal, während die Experimentalgruppe das gleiche OSA durchläuft, jedoch nach Abschluss des ersten (OSA 1) und vor Absolvierung des zweiten (OSA 2) den gesamten oder Teile des OV bearbeitet (Abb. 2.9). Die Operationalisierung der Nutzung des OV erfolgt über die Zugriffe (Klicks) auf die einzelnen Übungsaufgaben innerhalb des OV. Das OSA 1 enthält neben einem Fragebogen zu demographischen Variablen das fachspezifische OSA, welches das Vorwissen in den beiden

Abbildung 2.8
Ausschnitt Moodle Einführungsseite zum Themengebiet ›Kräfte‹ im OSA naturwissenschaftliche Grundlagen

Kräfte, Kraftvektor, Kräftesystem

Kräfte: Einführung

Im Rahmen der technischen Mechanik befassen wir uns mit Kräften und ihren Auswirkungen auf Körper. Ein einfaches Beispiel hierfür ist ein Ball, welcher mit seinem Eigengewicht der Schwerkraft bzw. der Erdbeschleunigung $\vec{g}$ ausgesetzt wird. Die Konsequenz ist, dass der Ball zu Boden fällt. Um den Ball nun in der ruhenden Lage zu halten, müssen wir der Schwerkraft entgegen wirken.

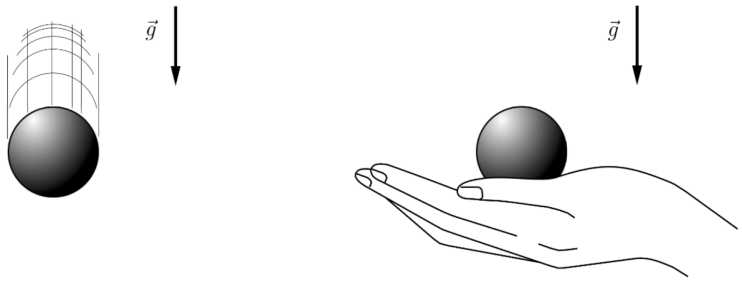


Abbildung 1: Ball im freien Fall (links) und gehaltener Ball (rechts). $\vec{g} = 9,81 \frac{m}{s^2}$ kennzeichnet die Erdbeschleunigung.

Die Kraft, die wir dabei aufbringen müssen, um den Ball in der ruhenden Lage zu behalten, ist definiert durch drei wesentliche Eigenschaften: **Betrag**, **Richtung** und **Angriffspunkt**. Diese Eigenschaften charakterisieren den Kraftbegriff in der Mechanik.

Der Kraftvektor

In der Mathematik wurde die Vektorrechnung bereits behandelt. Wir wissen, dass ein Vektor durch seine Richtung, seine Länge und seinen Ort im Koordinatensystem definiert ist. Diese Eigenschaften lassen sich auf eine auftretende Kraft übertragen.

⇒ **Kräfte werden mittels Vektoren beschrieben!**

1.) Der **Betrag** des Kraftvektors gibt an, wie groß die wirkende Kraft ist. Sie wird in der Einheit N (Newton) angegeben. In SI-Einheiten ist ein Newton folgendermaßen definiert:

$$N = \frac{kg \cdot m}{s^2}$$

Aus physikalischer Sicht bedeutet dies, dass ein 1 N die Kraft angibt, die einer Masse von 1 kg eine Beschleunigung von 1 $\frac{m}{s^2}$ erteilt. Aus mathematischer Sicht wissen wir, dass der Betrag eines Vektors dessen Länge angibt und folgendermaßen berechnet wird:

Für einen gegebenen Vektor $\vec{F} = \begin{pmatrix} F_x \\ F_y \\ F_z \end{pmatrix}$ ergibt sich der Betrag von $\vec{F}$ zu $\left|\vec{F}\right| = \sqrt{F_x^2 + F_y^2 + F_z^2}$.

Die Darstellung einer Kraft als Vektor führt uns zur zweiten, wesentlichen Eigenschaft einer Kraft: die Richtung.

2.) Die **Richtung** einer Kraft können wir durch ihre Wirkungslinie und den Richtungssinn auf der Wirkungslinie beschreiben. Die Wirkungslinie eines Kraftvektors stellt dar, auf welcher Linie der Kraftvektor unabhängig von seinem Angriffspunkt liegt (siehe Abbildung 2).

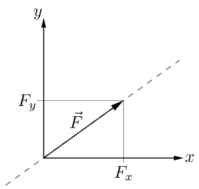
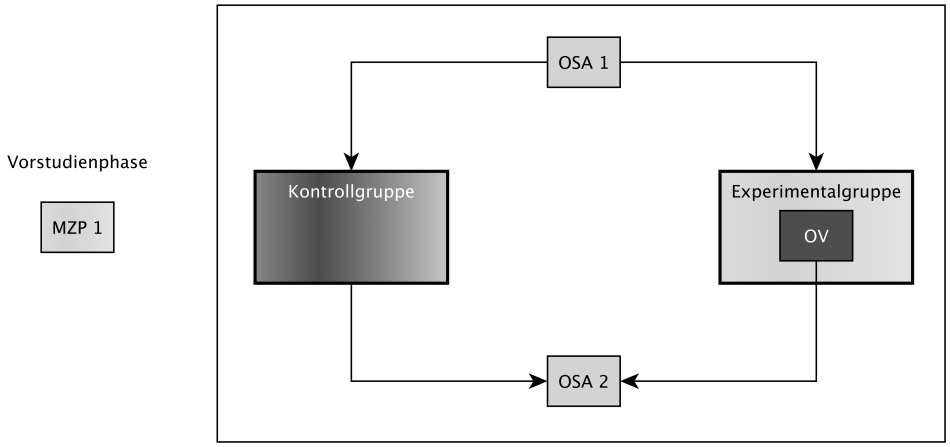


Abbildung 2: Vektor $\vec{F}$ mit Wirkungslinie (rot).

Inhaltsbereichen naturwissenschaftliche und mathematische Grundlagen mit entsprechenden Items testet. Im Gegensatz dazu wird im OSA 2 nur das fachspezifische OSA eingesetzt, die Items sind identisch mit denen des OSA 1. Die Effektivität des OV kann durch die mögliche Differenz der erreichten Punktzahlen in den beiden OSA bestimmt werden.

Abbildung 2.9

FUNDAMENT – Untersuchungsdesign Studienvorphase (in Anlehnung an Pelz et al., 2021, S. 103)



Anmerkungen. MZP = Messzeitpunkt; OSA = Online-Self-Assessment; OV = Online-Vorkurs.

Insgesamt konnten in der Studienvorphase (MZP 1) nur wenige Probanden für die Teilnahme gewonnen werden (Pelz et al., 2021). Probanden, die auch am OV teilgenommen haben, erzielten im zweiten OSA tendenziell bessere Ergebnisse als diejenigen, die den OV nicht absolviert haben (Pelz et al., 2021). Aufgrund der geringen Probandenzahl kann jedoch keine abschließende Aussage über die Wirksamkeit des OV getroffen werden, da insgesamt, insbesondere in der Kontrollgruppe die Probandenzahl zu gering ist (Pelz et al., 2021). Die Probandenzahl derjenigen, die den OV genutzt haben, liegt im einstelligen Bereich (Pelz et al., 2021). Somit sind auch hier keine generalisierbaren Aussagen möglich, jedoch kann tendenziell festgestellt werden, dass die Probanden, die sich mit dem OV auseinandergesetzt haben, diesen gewissenhaft und konsequent bearbeitet haben, da sie viele Aufgaben – im Durchschnitt bis zu 74 Aufgaben – bearbeitet haben (Pelz et al., 2021). Weitere

Details zu den erzielten Ergebnissen finden sich in Pelz et al. (2021).

Weiterführend untersucht die vorliegenden Arbeit, ob die Teilnahme bzw. die erreichten Ergebnisse im OSA 1 und/oder OSA 2 einen Einfluss auf den Studienerfolg (Klausurnoten) in den ersten beiden Semestern haben (Kap. 6.4.3).

2.5.1.2 FUNDAMENT - Studieneingangsphase

Um den Studienanfängern das Verständnis der Kernkonzepte der Technischen Mechanik zu erleichtern, wurden in FUNDAMENT für die Studieneingangsphase iOM entwickelt. Sie basieren auf einem ›Drei-Säulen-Konzept‹ und sind für die Veranstaltungen Technische Mechanik 1 und Technische Mechanik 2 konzipiert. Die erste der drei Säulen sind Lernvideos, die entweder als animierte Slideshows oder als Experimentvideos realisiert wurden. Darüber hinaus stehen den Studienanfängern mit JACK-Übungsaufgaben parametrisierte Online-Aufgaben mit automatischer Feedbackgenerierung zur Verfügung, sowie Online-Kommunikationsmöglichkeiten.

Experimentvideos sind Aufzeichnungen von Versuchen, die bereits in einer Vorlesung oder Übung präsentiert wurden (Abb. 2.10). Sie dienen der Nachbereitung im Selbststudium und sollen die Kernkonzepte der Technischen Mechanik veranschaulichen.

Der Schwerpunkt der animierten Slideshows liegt auf Rechenaufgaben, die in Hörsaalübungen und Tutorien behandelt werden (Abb. 2.11). Die animierten Slideshows vermitteln die konzeptionelle Herangehensweise sowie die Grundprinzipien der Technischen Mechanik bei der Bearbeitung entsprechender Aufgaben. Die Videos ergänzen das bestehende Lehr- und Lernangebot und ermöglichen eine zeit- und ortsunabhängige Wiederholung der Inhalte aus Vorlesungen, Hörsaalübungen und Tutorien.

Die zweite Säule der iOM sind die JACK-Übungsaufgaben, die in Moodle zur Verfügung gestellt werden. Wie die Lernvideos sind auch die JACK-Übungsaufgaben zeit- und ortsunabhängig und tragen somit zur Flexibilisierung des Studiums bei. Bei den JACK-Übungsaufgaben handelt es sich um parametrisierte Übungsaufgaben – jedes Öffnen einer Übungsaufgabe liefert den gleichen Aufgabenstamm, aber mit anderen

Abbildung 2.10
Ausschnitt eines Experimentvideos zum Thema ›Biegetheorie des dünnen prismatischen Balkens‹

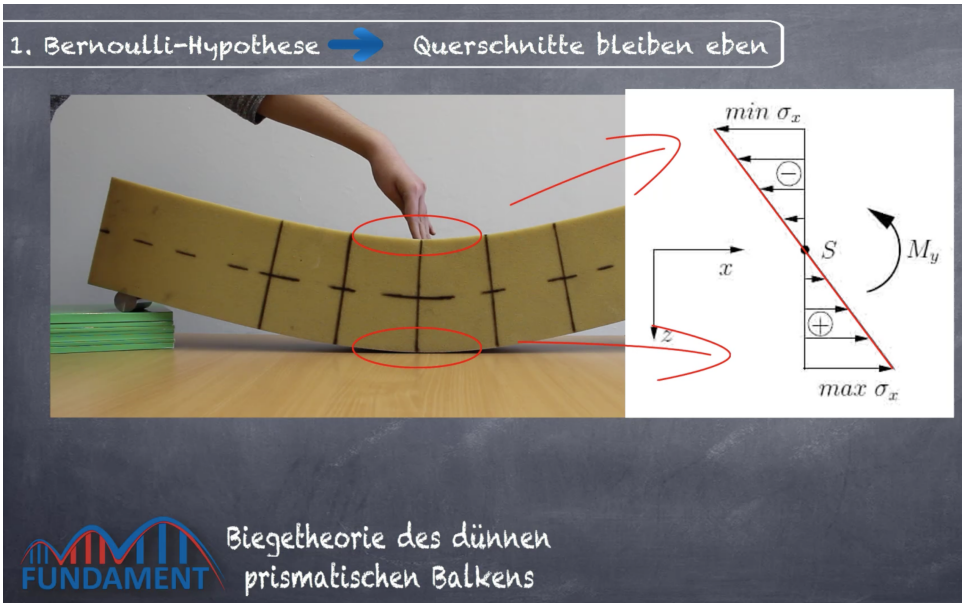
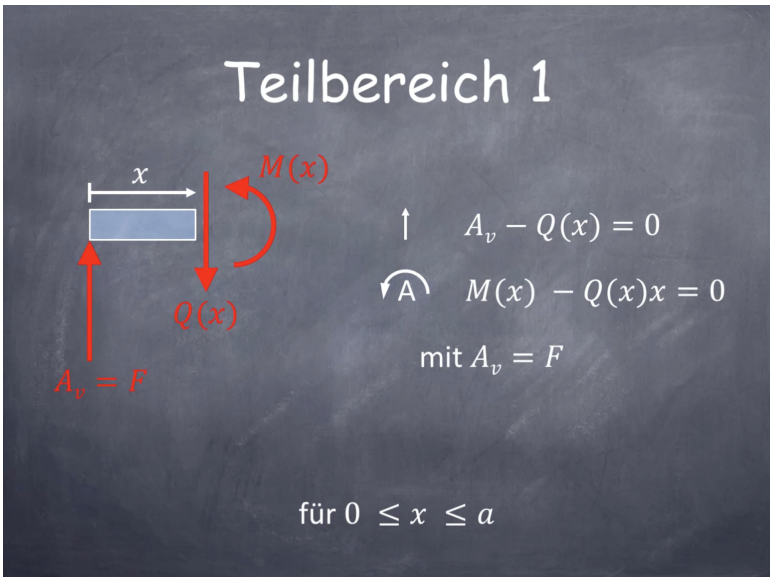


Abbildung 2.11
Ausschnitt einer animierten Slideshow zum Thema ›Schnittgrößen‹

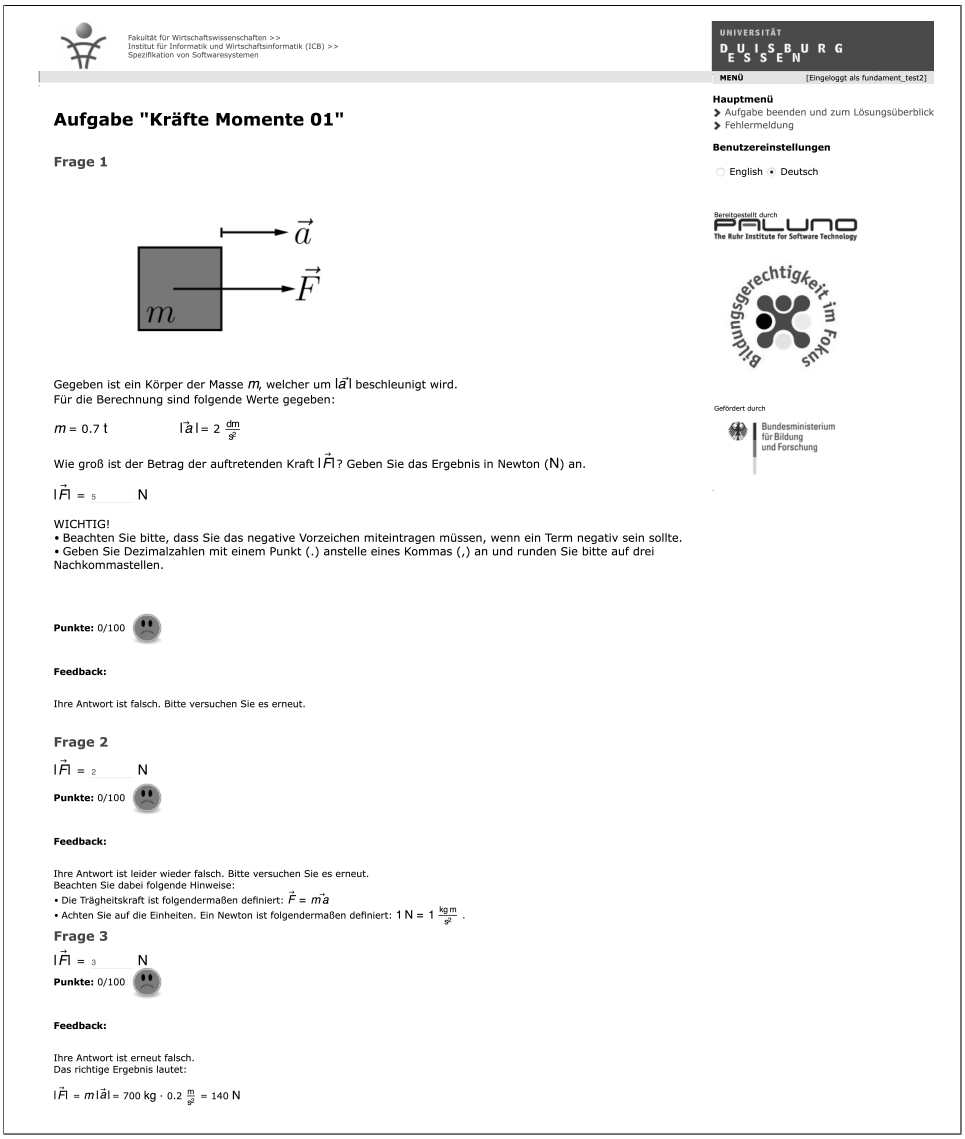


Zahlenwerten und entsprechend anderen Ergebnissen – in verschiedenen Schwierigkeitsgraden (Abb. 2.12). Das JACK-System ist in der Lage, automatisches Feedback zu den eingereichten Lösungen zu generieren und ggf. mehrstufige Hilfestellungen zu geben, um den Einstieg in eine Übungsaufgabe zu erleichtern. Darüber hinaus können typische Fehler mit entsprechenden Kommentaren hinterlegt werden, sodass Studierende explizit auf fehlerhafte Vorgehensweisen bei der Bearbeitung der Übungsaufgabe oder bei Teilergebnissen hingewiesen werden können.

Die Online-Kommunikation zwischen Studierenden und Lehrenden bzw. Studierenden untereinander stellt die dritte und damit letzte Säule der iOM dar. Zur Klärung organisatorischer und inhaltlicher Fragen wurden moderierte anonyme Moodle-Foren eingerichtet. Der dadurch generierte Austausch der Studierenden untereinander sollte zu einer stärkeren Integration der Studierenden in die akademische Gemeinschaft beitragen. Die Foren wurden jedoch von den Studierenden nicht angenommen, da sie alternative Kommunikationsmittel bevorzugen, weshalb sie in der Ergebnisdarstellung nicht weiter berücksichtigt werden (Pelz et al., 2020).

Auch in der Studieneingangsphase ist das Untersuchungsdesign nach dem klassischen Experimental-/Kontrollgruppendesign konzipiert Abb. 2.13). Während die Kontrollgruppe die Veranstaltungen Technische Mechanik 1 und Technische Mechanik 2 nach dem klassischen Vorlesungskonzept mit Hörsaalübungen und Tutorien besucht, nutzt die Experimentalgruppe zusätzlich die iOM. Die Überprüfung der Wirksamkeit erfolgt anhand der Ergebnisse der fachlichen Leistungstests in den Inhaltsbereichen Technische Mechanik und Rechenfähigkeit, die als Paper & Pencil-Tests an drei MZP durchgeführt werden: MZP 2: Beginn des ersten Fachsemesters, MZP 3: Ende des ersten Fachsemesters und MZP 4: Ende des zweiten Fachsemesters (weitere Details siehe Kap. 5.1). Zusätzlich werden die Klausurnoten der Technischen Mechanik (1. Fachsemester: TM 1/1.1/1.2; 2. Fachsemester: TM 2/2.1/ 2.2) und Mathematik berücksichtigt. Die Operationalisierung der Nutzung der iOM erfolgt über die Zugriffe (Klicks) auf die Videos und JACK-Übungsaufgaben.

Abbildung 2.12
Ausschnitt einer JACK-Übungsaufgabe zum Thema ›Kräfte und Momente‹



Fakultät für Wirtschaftswissenschaften >>
Institut für Informatik und Wirtschaftsinformatik (ICB) >>
Spezifikation von Softwaresystemen

UNIVERSITÄT DUISBURG ESSEN

MENÜ [Eingeloggt als fundament_test2]

Hauptmenü
> Aufgabe beenden und zum Lösungsüberblick
> Fehlermeldung

Benutzereinstellungen
English • Deutsch

Bereitgestellt durch PALUNO The Ruhr Institute for Software Technology

Bildungsgerechtigkeit im Fokus

Gefördert durch Bundesministerium für Bildung und Forschung

Aufgabe "Kräfte Momente 01"

Frage 1

Gegeben ist ein Körper der Masse m, welcher um $|\vec{a}|$ beschleunigt wird.
Für die Berechnung sind folgende Werte gegeben:

$m = 0.7$ t $\quad |\vec{a}| = 2 \frac{dm}{s^2}$

Wie groß ist der Betrag der auftretenden Kraft $|\vec{F}|$? Geben Sie das Ergebnis in Newton (N) an.

$|\vec{F}| =$ 5 N

WICHTIG!
• Beachten Sie bitte, dass Sie das negative Vorzeichen miteintragen müssen, wenn ein Term negativ sein sollte.
• Geben Sie Dezimalzahlen mit einem Punkt (.) anstelle eines Kommas (,) an und runden Sie bitte auf drei Nachkommastellen.

Punkte: 0/100

Feedback:

Ihre Antwort ist falsch. Bitte versuchen Sie es erneut.

Frage 2

$|\vec{F}| =$ 2 N

Punkte: 0/100

Feedback:

Ihre Antwort ist leider wieder falsch. Bitte versuchen Sie es erneut.
Beachten Sie dabei folgende Hinweise:
• Die Trägheitskraft ist folgendermaßen definiert: $\vec{F} = m\vec{a}$
• Achten Sie auf die Einheiten. Ein Newton ist folgendermaßen definiert: $1\,\text{N} = 1 \frac{kg\,m}{s^2}$.

Frage 3

$|\vec{F}| =$ 3 N

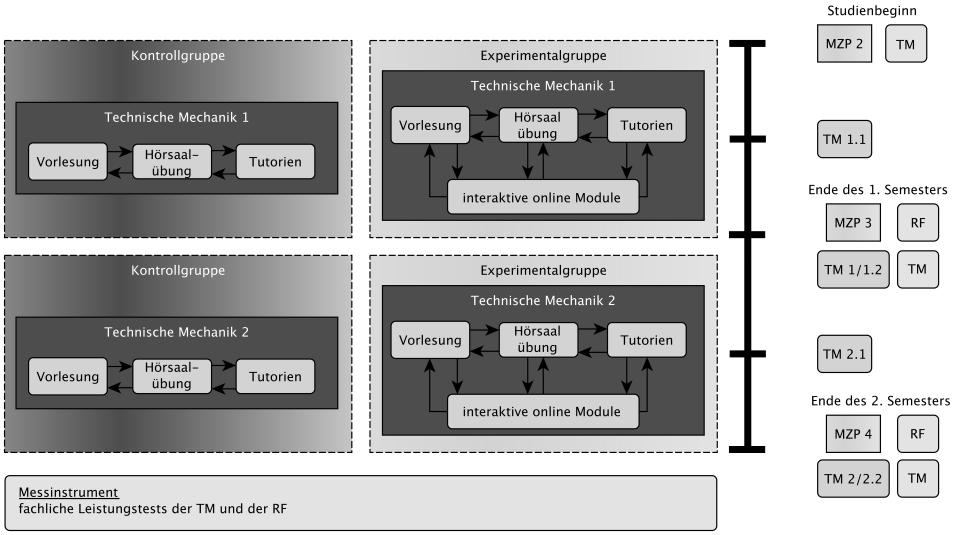
Punkte: 0/100

Feedback:

Ihre Antwort ist erneut falsch.
Das richtige Ergebnis lautet:

$|\vec{F}| = m|\vec{a}| = 700\,\text{kg} \cdot 0.2 \frac{m}{s^2} = 140\,\text{N}$

Abbildung 2.13
FUNDAMENT – Untersuchungsdesign Studieneingangsphase



Anmerkungen. MZP = Messzeitpunkt; TM = fachlicher Leistungstest der Technischen Mechanik; Rechenfähigkeit (RF) = Rechenfähigkeit; TM 1/1.1/1.2/2/2.1/2.2 = Klausuren ›Technische Mechanik‹.

Die Ergebnisse zeigen, dass die Personenfähigkeiten (IRT - Rasch-Modell [1pl]) über die drei MZP zunehmen, d.h. es ist ein Leistungszuwachs festzustellen (Pelz et al., 2020). Ebenso lassen sich signifikante Korrelationen zwischen den Personenfähigkeiten an den einzelnen MZP mit einem großen Effekt nachweisen (Pelz et al., 2020).

Die Betrachtung der Veranstaltungen der Technischen Mechanik erfolgt getrennt für die ersten beiden Fachsemester (1. Fachsemester: Technische Mechanik 1 bzw. 2. Fachsemester: Technische Mechanik 2). Die Berücksichtigung der iOM erweist sich jedoch – wie schon bei MZP 1 – als schwierig, da die Kontrollgruppe nicht die notwendige Größe aufweist. Die Differenzierung der Studierenden erfolgt in Nutzergruppen (Nutzung von Videos und JACK-Übungsaufgaben, nur Videos, nur JACK-Übungsaufgaben, keine iOM-Nutzung). Dabei bilden die Studierenden ohne iOM-Nutzung die Kontrollgruppe, deren Größe bei beiden Fachsemestern nur knapp 10% der Stichprobe beträgt (Pelz et al., 2020). Die Kontrollgruppe verkleinert sich weiter, wenn die entsprechenden Personenfähigkeiten betrachtet werden sollen, da

die relevanten Studierenden nicht an den eingesetzten fachlichen Leistungstests teilgenommen haben und somit keine Personenfähigkeiten ermittelt werden können. Aussagen über die Wirksamkeit der iOM sind daher nur eingeschränkt möglich und eher als Tendenzen zu verstehen. So zeigt sich, dass signifikante Zusammenhänge zwischen der Nutzung der JACK-Übungsaufgaben und den entsprechenden Klausurnoten der Technischen Mechanik bestehen (Pelz et al., 2020). Dagegen hat die Nutzung der Videos keinen signifikanten Effekt auf die Klausurnoten der Technischen Mechanik (Pelz et al., 2020). Insgesamt werden die iOM von den Studierenden gut angenommen, wie die durchaus positiven Nutzungszahlen zeigen, wobei die Nutzung der iOM im ersten Fachsemester höher ist als im zweiten (Pelz et al., 2020).

Ob die Nutzung der iOM einen Einfluss auf die Einstufung in ein Kompetenzniveau hat, wird in dieser Arbeit untersucht.

2.5.2 ALSTER

Die DFG-Forschergruppe *Akademisches Lernen und Studienerfolg in der Eingangsphase von naturwissenschaftlich-technischen Studiengängen* (ALSTER)[3] untersucht in mehreren Teilprojekten, auf welche Faktoren der Studienerfolg in den Studiengängen der Mathematik, Informatik, Naturwissenschaften und Technik (MINT) zurückzuführen ist. Beschreibungen und Ergebnisse der fächerübergreifenden zentralen Datenerhebung finden sich u.a. in Fleischer et al. (2017, 2019). Der ingenieurwissenschaftliche Kontext wird in der Forschergruppe in ›Teilprojekt C: Studienerfolg in der Physik und im Bauingenieurwesen unter besonderer Berücksichtigung der geforderten und benötigten mathematischen Kompetenz‹ aufgegriffen. Da der Bereich der Physik im Kontext dieser Arbeit weniger relevant ist, impliziert die Bezeichnung ALSTER im Folgenden den Teil des Teilprojekts C, der sich auf das Studium des Bauingenieurwesen bezieht. Die Datenerhebung erfolgte an zwei Standorten im Ruhrgebiet, der UDE und der RUB.

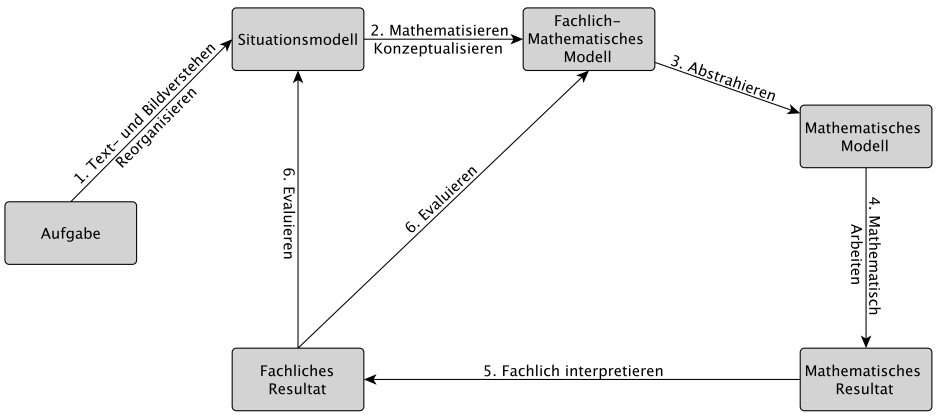
ALSTER baut auf dem Konstrukt der mechanisch-mathematischen Modellierungsfähigkeit auf, dessen Grundlage das Konstrukt des mathematischen Modellierens ist, welches in verschiedenen Abwandlungen

[3]FOR 2242, Projektnummer 257652630.

u.a. von Borromeo Ferri (2010) und Borromeo Ferri et al. (2013) in der Fachdidaktik Mathematik vorgeschlagen wird (Dammann & Lang, 2018). Die Verknüpfung von realen Situationen mit mathematischen Prinzipien/Methoden, mit denen spezifische Fragen zu diesen konkreten Sachverhalten beantwortet werden können, ist der zentrale Aspekt des Modellierung (Dammann & Lang, 2018). Die dazu notwendigen Modellierungsprozesse sind zumeist als Kreisläufe im Sinne idealisierter Prozessbeschreibungen zu verstehen, abweichend davon werden in der realen Bearbeitung einzelne Prozessschritte vertauscht oder mehrfach durchlaufen (Dammann & Lang, 2018). Borromeo Ferri et al. (2013) definieren diese Fähigkeit, bei der »die jeweils nötigen Prozessschritte beim Hin- und Herwechseln zwischen Realität und Mathematik problemadäquat auszuführen sowie gegebene Modelle zu analysieren oder vergleichend zu beurteilen« (Borromeo Ferri et al., 2013, S. 18), als Modellierungskompetenz.

In ALSTER wurde ein Modell zur fachlich-mathematischen Modellierung (Abb. 2.14) entwickelt (Dammann & Lang, 2018). Im Unterschied zu anderen Modellen wird mit dem fachlich-mathematischen Modell eine weitere Modellierungsphase in den Modellierungskreislauf eingeführt, die die Verwendung naturwissenschaftlich-technischer Symbole berücksichtigt (Dammann & Lang, 2018).

Abbildung 2.14
ALSTER – Modell fachlich-mathematischer Modellierung (Dammann & Lang, 2018, 145, eigene Darstellung)



Der Modellierungsprozess (Abb. 2.14) wird auf eine typische Aufgabenstellung angewendet, die Dammann und Lang (2018) als eine vordefinierte Situation beschreiben, die bereits eine Abstraktion durchlaufen hat, um das Verständnis der Aufgabe und die Anwendung der zugehörigen Lösungsverfahren zu erleichtern. Aus der Aufgabenstellung wird das Situationsmodell entwickelt, aus dem durch fachspezifische Konzepte das fachlich-mathematische Modell entsteht, welches wiederum durch Abstraktion in ein vollständig mathematisches Modell überführt wird (Dammann & Lang, 2018). Aus dem mathematischen Modell wird durch mathematisches Arbeiten (Anwendung mathematischer Prinzipien/Methoden) ein mathematisches Resultat, welches wie bereits im zweiten Modellierungsschritt wieder auf die fachlichen Konzepte zurückgreift, fachlich interpretiert und somit in ein fachliches Resultat überführt wird (Dammann & Lang, 2018). Das fachliche Resultat kann anschließend mit dem Situationsmodell und/oder dem fachlich-mathematischen Modell evaluiert werden (Dammann & Lang, 2018).

In ALSTER wurde für das Studium des Bauingenieurwesens das Fach Technische Mechanik fokussiert, in dem die Modellierungsfähigkeit bereits in den ersten Fachsemestern eine zentrale Rolle spielt (Dammann & Lang, 2018). Die Bearbeitungsschritte typischer Aufgabenstellungen der Technischen Mechanik folgen dem Modell des fachlich-mathematischen Modellierens, wobei für das Bauingenieurwesen modifiziert von der mechanisch-mathematischen Modellierungsfähigkeit gesprochen wird (Dammann & Lang, 2018). Die sechs typischen Bearbeitungsschritte von Aufgaben der Technischen Mechanik, die Magnus und Müller-Slany (2009) in ihrem Lehrbuch benennen, dienen in ALSTER als Grundlage. Aus heutiger Sicht hat Müller-Slany (2018) diese Punkte zu einer allgemeinen Systematik der Modellbildung erweitert, die für die Lösungsfindung in der Technischen Mechanik geeigneet ist und die reale technische Systeme zum Ausgangspunkt des Modellierungsprozesses macht. Als »übergreifende Kompetenz« (Dammann und Lang, 2018, S. 145) wird in ALSTER die mechanisch-mathematische Kompetenz verstanden (Dammann & Lang, 2018). In Anlehnung an Weinert (2014) (siehe Kap. 2.1) setzt die Modellierungskompetenz fachliches, fachlich-

mathematisches und mathematisches Wissen sowie die Fähigkeit zu dessen Anwendung voraus (Dammann & Lang, 2018). Zwischen den drei Dimensionen Fachwissen, Modellierungsfähigkeit und Rechenfähigkeit zeigen sich mittlere signifikante Zusammenhänge, sie weisen also eine Nähe zueinander auf, sind aber dennoch empirisch differenzierbar (Dammann & Lang, 2018). Ein Beitrag zu den theoretischen Zusammenhängen von Fachwissen, Modellierungsfähigkeit und Rechenfähigkeit ist bei Dammann in Arbeit.

Das Untersuchungsdesign von ALSTER sieht drei MZP vor: MZP 1 - Beginn des ersten Fachsemesters, MZP 2 - Ende des ersten Fachsemesters und MZP 3 - Ende des zweiten Fachsemesters (Dammann & Lang, 2019). An jedem MZP werden die drei Konstrukte mit separaten Testinstrumenten erfasst: Fachwissen der Mechanik (FW), mechanisch-mathematische Modellierungsfähigkeit (MF) und grundlegende Rechenfähigkeit (RF) (Dammann & Lang, 2018).

Bei der Entwicklung des Testinstruments zum Fachwissen wurden die als empirisch belastbar geltenden Facetten der Komplexitätsdimension (vgl. u.a. Kauertz, 2008) berücksichtigt (Dammann & Lang, 2018). Das Testinstrument zur Modellierungsfähigkeit überprüft, ob die Probanden über Modellierungskompetenz verfügen und somit in der Lage sind, die einzelnen Modellierungsschritte innerhalb des mechanisch-mathematischen Modellierungskreislaufs zu bewältigen (Dammann & Lang, 2018). Die Items sind so konstruiert, dass der Itemstamm eine Modellierungsphase und die Handlungsempfehlung einen Modellierungsschritt zur nächsten Modellierungsphase darstellt (Dammann & Lang, 2018). Die Items bilden jedoch nur die Modellierungsschritte zwei, drei und fünf ab, da die Schritte eins und sechs im ersten Studienjahr nicht explizit vermittelt werden und der Schritt vier (mathematisches Arbeiten) bereits durch das Testinstrument zur Rechenfähigkeit erfasst wird (Dammann & Lang, 2018). Der Lehrplan der Sekundarstufe II wird im Testinstrument zur Rechenfähigkeit mit den Inhaltsbereichen Rechnen, Terme und Gleichungen, Trigonometrie, Differenzieren und Integrieren abgebildet, ist aber fächerübergreifend und nicht fachspezifisch konzipiert (Dammann & Lang, 2018).

Neben der beschriebenen Herangehensweise werden Operatoren be-

nötigt, die das Vorgehen der Probanden innerhalb der einzelnen Modellierungsschritte beschreiben, wobei die bereits erwähnten Schritte zur Lösung mechanischer Aufgaben von Magnus und Müller-Slany (2009) berücksichtigt werden (Dammann & Lang, 2018). Die den drei relevanten Modellierungsschritten zugeordneten Operatoren sind:

- 2. Modellierungsschritt (Mathematisieren/Konzeptualisieren)
  - Freischneiden der Kräfte eines mechanischen Systems
  - Benennung fachlicher Konzepte und Prinzipien
- 3. Modellierungsschritt (Abstrahieren)
  - Benennung fachlich-mathematischer Konzepte und Prinzipien
  - Aufstellen der fachlich-mathematischen Gleichungen
- 5. Modellierungsschritt (fachlich Interpretieren)
  - Benennen (fachlich) mathematischer Resultate
  - Reflexion (fachlich) mathematischer Resultate (Dammann & Lang, 2018).

MZP 1 liegt zu Beginn des Studiums und dementsprechend müssen die Testinstrumente Inhalte des Schulcurriculums abdecken, dies ist in Mathematik unproblematisch, da es aber kein Schulfach Technische Mechanik gibt, beziehen sich die Items auf den Bereich Mechanik der Schulphysik (Dammann & Lang, 2018).

Die Testinstrumente wurden an allen MZP als Paper & Pencil-Tests eingesetzt, zusätzlich wurden zu Studienbeginn demographische Variablen wie u.a. die Hochschulzugangsberechtigung oder die kognitive Grundfähigkeit erhoben (Dammann & Lang, 2019). Der Studienerfolg wird über die Klausurnoten der Technischen Mechanik und Mathematik operationalisiert (Dammann & Lang, 2019).

Dammann und Lang (2019) können zeigen, dass die fachlichen Leistungen (Technische Mechanik und Mathematik) in dem in ALSTER entwickelten und eingesetzten Testinstrument zu Beginn des ersten

Fachsemesters mit den Leistungen in den Technische Mechanik- bzw. Mathematikklausuren und damit mit dem Studienerfolg korrelieren. Es stellt sich heraus, dass nicht die Inhaltsbereiche der Mathematikitems entscheidend sind, sondern schwierigkeitsbestimmende Merkmale, die allerdings in ALSTER nicht näher untersucht wurden (Dammann & Lang, 2019).

In einer weiteren Untersuchung der fachlichen Leistung der Mathematik konnten J. Müller et al. (2018) belegen, das ein signifikanter Zusammenhang zwischen mathematischem Wissen und Studienerfolg besteht. Dies gilt auch, wenn für die Abiturnote oder die schulische Mathematiknote in der Betrachtung kontrolliert werden (J. Müller et al., 2018). Das mathematische Wissen klärt auch die substanzielle inkrementelle Varianz auf und ist signifikant gegenüber der Abiturnote und der Mathematiknote in der Schule (J. Müller et al., 2018).

Ob sich die in ALSTER erhobenen Befunde hinsichtlich der fachlichen Leistungen replizieren lassen (FUNDAMENT), wird in dieser Arbeit untersucht (Kap. 6.4.1).

Kapitel 3

Forschungsfragen und Hypothesen

Forschungsfragen (F) und Hypothesen (H) sind in der empirischen Forschung notwendig, um ein Forschungsproblem einzugrenzen (Döring & Bortz, 2016). Diese lassen sich aus den theoretischen Grundlagen ableiten, die in Kap. 2 dargestellt sind.

Die Daten der Studien FUNDAMENT und ALSTER wurden zu unterschiedlichen Zeitpunkten erhoben (siehe Kap. 5.1). Daher muss untersucht werden, ob es signifikante Unterschiede zwischen den Ergebnissen der beiden Studien gibt. Dies ist in F1 formuliert:

> **F1** Können die Befunde der Studie ALSTER hinsichtlich der Inhaltsbereiche Technische Mechanik und Rechenfähigkeit in der Studie FUNDAMENT repliziert werden?

In Bezug auf die erste Forschungsfrage (F1) können folgende Hypothesen aufgestellt werden:

- **H1.1** Die erreichten Personenfähigkeiten in dem Inhaltsbereich Technische Mechanik an den MZP 2 bis 4 zeigen keine signifikanten Unterschiede zwischen den beiden Studien (FUNDAMENT und ALSTER).

- **H1.2** Die unterschiedlichen Antwortformate (FUNDAMENT: geschlossen, ALSTER: offen) der Testinstrumente der Rechenfähigkeit haben keinen Einfluss auf die Rangfolge der Itemschwierigkeiten, so dass eine gemeinsame Skalierung möglich ist.

Die separate Niveaumodellierung der beiden Inhaltsbereiche (Technische Mechanik und Rechenfähigkeit) kann im Zusammenhang mit dem Studienerfolg in der Studieneingangsphase untersucht werden. Unter Studienerfolg wird die erreichte Klausurnote (Technische Mechanik bzw. Mathematik) am Ende des jeweiligen Fachsemesters verstanden (siehe Kap. 5.1). Daraus resultiert die zweite Forschungsfrage (F2):

> **F2** Gibt es signifikante Zusammenhänge zwischen den modellierten Kompetenzniveaus (Technische Mechanik bzw. Rechenfähigkeit) und dem Studienerfolg, sodass die Niveaus als Prädiktoren für den Studienerfolg angesehen werden können?

Die formulierten Annahmen für die zweite Forschungsfrage (F2) lauten:

- **H2.1** Das erreichte Kompetenzniveau (Technische Mechanik) an den MZP 2 und 3 korreliert negativ mit dem Studienerfolg nach dem ersten Fachsemester (›Technische Mechanik 1‹).
- **H2.2** Das erreichte Kompetenzniveau (Technische Mechanik) an den MZP 3 und 4 korreliert negativ mit dem Studienerfolg nach dem zweiten Fachsemester (›Technische Mechanik 2‹).
- **H2.3** Das erreichte Kompetenzniveau (Rechenfähigkeit) an den MZP 2 und 3 korreliert negativ mit dem Studienerfolg nach dem ersten Fachsemester (›Mathematik 1‹).
- **H2.4** Das erreichte Kompetenzniveau (Rechenfähigkeit) an dem MZP 3 und 4 (FUNDAMENT) korreliert negativ mit dem Studienerfolg nach dem zweiten Fachsemester (›Mathematik 2‹).

Die in der Studie FUNDAMENT entwickelten Online-Elemente (OSA und iOM) können einen Einfluss auf den Studienerfolg haben. Daraus ergibt sich für die Studienvorphase die dritte Forschungsfrage (F3):

> **F3** Haben die Teilnahme und die entsprechenden Ergebnisse am OSA an MZP 1 (FUNDAMENT) einen Einfluss auf den Studienerfolg?

Die Hypothesen zur dritten Forschungsfrage (F3) lauten:

- **H3.1** Die Teilnahme am OSA naturwissenschaftlicher oder mathematischer Grundlagen führt zu signifikant besseren Technische Mechanik-Klausurnoten (›Technische Mechanik 1‹ bzw. ›Technische Mechanik 2‹) mit geringer Effektstärke.
- **H3.2** Die Teilnahme am OSA naturwissenschaftlicher oder mathematischer Grundlagen führt zu signifikant besseren Mathematik-Klausurnoten (›Mathematik 1‹ bzw. ›Mathematik 2‹) mit geringer Effektstärke.
- **H3.3** Die erreichten Summenscores im OSA naturwissenschaftlicher bzw. mathematischer Grundlagen korrelieren schwach negativ mit dem Studienerfolg am Ende des ersten (›Technische Mechanik 1‹ bzw. ›Mathematik 1‹) bzw. zweiten Fachsemester (›Technische Mechanik 2‹ bzw. ›Mathematik 2‹).

In den ersten beiden Fachsemestern werden im Rahmen der Studie FUNDAMENT für den Inhaltsbereich Technische Mechanik iOM (JACK-Übungsaufgaben und Videos) eingesetzt, die sich auf das Erreichen eines Kompetenzniveaus auswirken können. Für die ersten beiden Fachsemester lässt sich somit die vierte Forschungsfrage formulieren (F4):

> **F4** Besteht ein Zusammenhang zwischen den modellierten Kompetenzniveaus (Technische Mechanik) an den MZP 2-4 und der Nutzung der iOM im ersten bzw. zweiten Fachsemester?

Zur vierten Forschungsfrage (F4) können folgende Hypothesen formuliert werden:

- **H4.1** Das erreichte Kompetenzniveau (Technische Mechanik) an den MZP 2 und 3 korreliert mit der Nutzung der iOM im ersten Fachsemester.
- **H4.2** Das erreichte Kompetenzniveau (Technische Mechanik) an den MZP 3 und 4 korreliert mit der Nutzung der iOM im zweiten Fachsemester.

Um die genannten Forschungsfragen beantworten zu können, muss eine Niveaumodellierung durchgeführt werden. Dazu werden zunächst IRT skalierte Daten benötigt. Die entsprechenden methodischen Grundlagen zur IRT und Niveaumodellierung werden im folgenden Kapitel gelegt.

Kapitel 4

Methodik

Auf der Grundlage der im Rahmen der Studien FUNDAMENT und ALSTER erhobenen Daten sollen Kompetenzniveaumodelle für die Technische Mechanik und die Rechenfähigkeit entwickelt werden. Die dafür notwendigen methodischen Grundlagen werden in diesem Kapitel gelegt.

Die erhobenen Daten sollen nach der probabilistischen Testtheorie skaliert werden. Dementsprechend werden Modelle der IRT eingeführt. Ein besonderer Schwerpunkt liegt dabei auf dem 1pl-Rasch-Modell, aber auch die 2pl/3pl-Birnbaum-Modelle werden präsentiert sowie Methoden, die einen Modellvergleich ermöglichen. Relevante Schätzverfahren zur Bestimmung von Itemschwierigkeiten und Personenfähigkeiten werden ebenfalls vorgestellt.

Die methodischen Grundlagen der IRT sind Voraussetzung für das Verständnis der Kompetenzniveaumodellierung. Entsprechende Verfahren zur Bestimmung von Kompetenzniveaus werden erläutert, wobei der Fokus auf der Post-hoc-Kompetenzniveaumodellierung nach Beaton und Allen (1992) liegt.

4.1 Item-Response-Theorie

Manifeste Variablen, die in den Testinstrumenten der fachlichen Leistungstests erhoben werden, sollen Rückschlüsse auf latente Variablen ermöglichen. Als manifeste Variablen werden beobachtete Antworten

bezeichnet, wie z.B. die Antworten eines Probanden auf ein Test-Item (Döring & Bortz, 2016). Latente Variablen hingegen sind definiert als »nicht direkt beobachtbare... Konstrukte« (Döring und Bortz, 2016, S. 483), die ihrerseits auf manifeste Variablen zurückgeführt werden können. Testtheorien eignen sich zur Beschreibung der Zusammenhänge zwischen manifesten und latenten Variablen (Döring & Bortz, 2016). Die Klassische Testtheorie (KTT) und die IRT (auch probabilistische Testtheorie genannt [Döring und Bortz, 2016]) sind die beiden gebräuchlichsten Testtheorien zur Interpretation von Testwerten (Moosbrugger et al., 2020).

Die KTT geht von der Annahme aus, dass sich ein Testergebnis aus der »wahren Merkmalsausprägung ... [und] einer den Testwert vergrößernden oder verkleinernden Fehlerkomponente« (Döring und Bortz, 2016, S. 461) zusammensetzt. Es sind also sowohl der Testwert als auch die Fehlerkomponente erforderlich, damit die KTT aussagekräftige Ergebnisse liefert (Döring & Bortz, 2016). Bei diesem deterministischen Ansatz entspricht die Merkmalsausprägung (und die Fehlerkomponente) direkt dem Testergebnis (Döring & Bortz, 2016). Entgegen der früheren Annahme gibt es mittlerweile Ansätze, mit denen nicht nur die typischerweise intervallskalierten Itemvariablen, sondern auch kontinuierliche oder geordnete kategoriale Itemvariablen analysiert werden können (Moosbrugger et al., 2020).

Beim probabilistischen Ansatz der IRT wird die Merkmalsausprägung nicht direkt aus dem Testergebnis abgeleitet, sondern es wird die Merkmalsausprägung, »die für verschiedene Arten der Item-Beantwortung am wahrscheinlichsten« (Döring und Bortz, 2016, S. 461) ist, geschätzt. Der IRT liegt die Annahme zugrunde, dass das Merkmal eines Items (Itemschwierigkeitsparameter[4]) und das latente Personenmerkmal (Personenfähigkeitsparameter[5]) die Wahrscheinlichkeit eine bestimmte Antwort zu geben bedingen (Döring & Bortz, 2016). Diese Wahrscheinlichkeit kann durch eine Wahrscheinlichkeitsfunktion beschrieben werden (Moosbrugger et al., 2020). Üblicherweise werden kategoriale Daten mit der IRT skaliert, aber es gibt bereits Erweiterungen, die bestehende An-

[4]Nachfolgend verkürzt als Itemschwierigkeit bezeichnet.
[5]Nachfolgend verkürzt als Personenfähigkeit bezeichnet.

sätze auf kontinuierliche Itemvariablen ausdehnen (Moosbrugger et al., 2020). Die IRT wurde hauptsächlich für dichotome Testitems entwickelt, d.h. Items, die zwei Ausprägungen haben (z.B. richtig = ›1‹ und falsch = ›0‹) (Moosbrugger et al., 2020). In der Literatur werden keine genauen Grenzen für den Stichprobenumfang bei der IRT-Skalierung angegeben. Koller et al. (2015) bezeichnen eine Stichprobe mit $n = 30$ und fünf Items jedoch als sehr klein. Eine entsprechende Skalierung bei einer solchen Stichprobengröße ist daher nicht empfehlenswert.

Die IRT hat sich in der empirischen Bildungsforschung etabliert und wird bspw. zur Kompetenzdiagnostik in der schulischen Bildung (siehe Kap. 2.1.2) bei TIMSS oder PISA, aber auch im Hochschulbereich (siehe Kap. 2.1.2) eingesetzt.

Hieran orientiert sich auch diese Arbeit. Da die fachlichen Leistungstests ausschließlich Testinstrumente mit dichotomen Merkmalsausprägungen besitzen, bietet sich eine Auswertung im Sinne der IRT an. Deswegen werden in den folgenden Kapiteln verschiedene Methoden der IRT, wie auch Eigenschaften und Möglichkeiten der Modellprüfung erläutert.

4.1.1 1pl-Modell nach Rasch

Das einparametrische logistische Modell (1pl), auch (dichotomes) Rasch-Modell genannt, ist das gebräuchlichste Latent-Trait-Modell und wird hauptsächlich in der Leistungsdiagnostik zur Bestimmung der Leistungsfähigkeit eingesetzt (Kelava & Moosbrugger, 2020). Latent-Trait-Modelle setzen voraus, dass eine zugrunde liegende kontinuierliche latente Eigenschaft einer Person (Personenfähigkeit) ein Verhalten (Antwort auf ein Item) hervorruft (Kelava & Moosbrugger, 2020). Diese Personenfähigkeit kann bei verschiedenen Personen unterschiedlich ausgeprägt sein (Gollwitzer, 2020). Im Gegensatz dazu werden bei Latent-Class-Modellen keine individuellen Personenfähigkeiten (quantitative kontinuierliche Werte) geschätzt, sondern eine Person wird aufgrund ihres Antwortverhaltens im betrachteten Testinstrument einer latenten Klasse zugeordnet (Kelava & Moosbrugger, 2020). Im Rahmen dieser Arbeit liegt der Schwerpunkt auf Latent-Trait-Modellen, weitere Details zu Latent-Class-Modellen finden sich u.a. in Gollwitzer (2020, Kap. 22)

und Rost (2004, Kap. 3.1.2.2).

Die Beschreibung dichotomer Daten im Rahmen des Rasch-Modells erfolgt durch folgende Modellgleichung (Strobl, 2012):

$$P\left(U_{ij} = u_{ij} \mid \theta_i, \beta_j\right) = \frac{e^{u_{ij}(\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}} \tag{4.1}$$

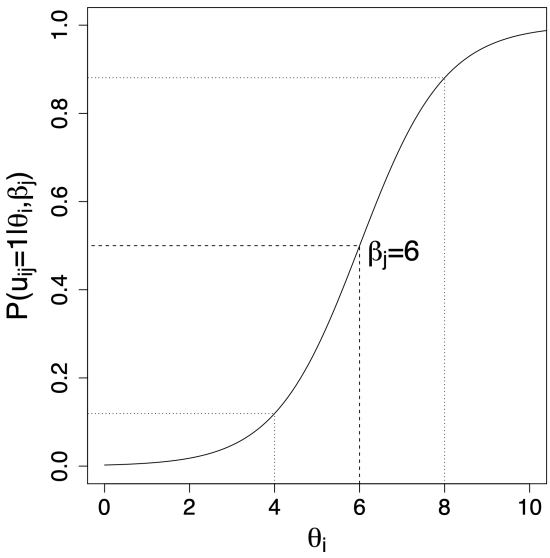
- θ_i = Personenfähigkeit der Person i $(i = 1, ..., n)$
- β_j = Itemschwierigkeit des Items j $(j = 1, ..., m)$
- U_{ij} = Itemvariable (Antwort der Person i auf das Item j)
- u_{ij} = konkreter Wert der Variablen U_{ij} (0 oder 1) (Strobl, 2012).

Diese Modellgleichung (Gl. 4.1) berechnet die Wahrscheinlichkeit $P\left(U_{ij} = u_{ij} \mid \theta_i, \beta_j\right)$, mit der eine Person i mit der Personenfähigkeit θ_i auf ein Item j mit der Itemschwierigkeit β_j antwortet (u_{ij}) (Strobl, 2012).

In der Literatur finden sich auch andere Schreibweisen dieser Modellgleichung, wobei meist $u_{ij} = 1$ verwendet wird, d.h. bei dichotomen Items wird die Antwort ›1‹ für eine richtige Antwort festgelegt (u.a. Kelava & Moosbrugger, 2020; Rost, 2004). Teilweise werden auch andere griechische Buchstaben für die Personenfähigkeit und die Itemschwierigkeit verwendet (u.a. Kelava & Moosbrugger, 2020; Rost, 2004). Kelava und Moosbrugger (2020) bezeichnen diese Gleichung als Itemcharakterische Funktion und weisen auf die Monotonie dieser Funktion hin. Durch die Monotonie der Funktion implizieren »höhere latente Merkmalsausprägungen ... [(höhere Personenfähigkeiten)] eine höhere Lösungswahrscheinlichkeit« (Kelava und Moosbrugger, 2020, S. 373). Dies wird auch an der rechten Seite der Modellgleichung (Gl. 4.1) deutlich: In den Exponentialfunktionen finden sich jeweils die Differenz aus Personenfähigkeit und Itemschwierigkeit $(\theta_i - \beta_j)$, somit hängt die Lösungswahrscheinlichkeit eines Items von beiden Parametern ab (Strobl, 2012). Wird bei einer richtigen Antwort $u_{ij} = 1$ gesetzt, kann die Modellgleichung umgeformt und als logistische Funktion bezeichnet werden (Strobl, 2012):

$$\frac{e^x}{1 + e^x} \tag{4.2}$$

Abbildung 4.1
Itemchararakteristikkurve für ein Item (Itemschwierigkeit $\beta = 1.1$) (Strobl, 2012, S. 10)

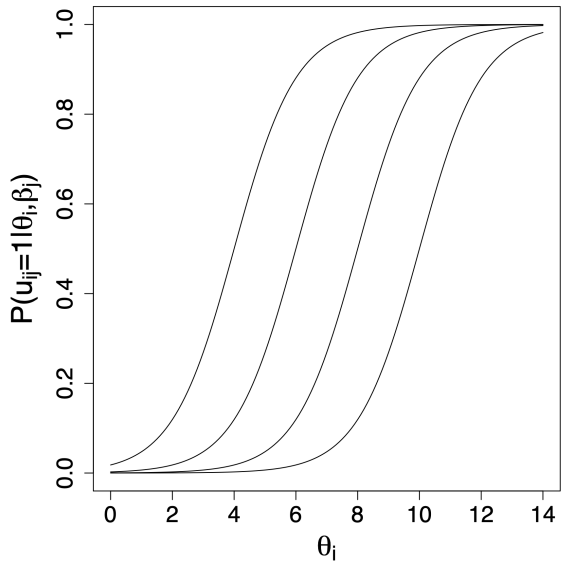


Anmerkungen. $P(U_{ij} = 1 \mid \theta_i, \beta_j)$ = Lösungswahrscheinlichkeit; θ_i = Personenfähigkeit der Person i; β_j = Itemschwierigkeit des Items j.

Abgeleitet vom Funktionsverlauf der logistischen Funktion (Gl. 4.2) kann der Funktionsverlauf der Modellgleichung (Gl. 4.1) als sogenannte Itemchararakteristikkurve (ICC) (engl.: *item characteristic curve*) dargestellt werden (Abb. 4.1). Der Verlauf der ICC gibt die Lösungswahrscheinlichkeit eines Items mit einer bestimmten Itemschwierigkeit in Abhängigkeit von einer Person mit einer bestimmten Personenfähigkeit an (Kelava & Moosbrugger, 2020; Koller et al., 2012; Rost, 2004; Strobl, 2012). Auf der x-Achse sind sowohl die Personenfähigkeit als auch die Itemschwierigkeit auf einer gemeinsamen Skala (engl.: *Joint Scale*) (Kelava & Moosbrugger, 2020) in einem Wertebereich von $-\infty$ bis $+\infty$ aufgetragen (Koller et al., 2012). Auf der y-Achse ist dagegen der Wertebereich von 0 bis 1 dargestellt, der die Lösungswahrscheinlichkeit eines Items angibt (Koller et al., 2012). Die Itemschwierigkeit eines Items ist definiert als der Punkt, an dem die Lösungswahrscheinlichkeit $P(1 \mid \theta_i, \beta_j) = 50\%$ beträgt, dies ist der Wendepunkt der ICC (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012).

In Abb. 4.1 wird eine ICC durch ein Item mit der Itemschwierigkeit

Abbildung 4.2
Itemchararakteristikkurve für mehrere Items mit unterschiedlichen Itemschwierigkeiten (Strobl, 2012, S. 11)



Anmerkungen. $P(U_{ij} = 1 \mid \theta_i, \beta_j)$ = Lösungswahrscheinlichkeit; θ_i = Personenfähigkeit der Person i.

$\beta = 6$ repräsentiert. Aufgrund der gemeinsamen Skala kann man an der gestrichelten Linie ablesen, dass eine Person mit der gleichen Personenfähigkeit ($\theta = 6$) dieses Item mit einer Wahrscheinlichkeit von 50% lösen kann. Andere Personen mit einer höheren Personenfähigkeit würden dieses Item ebenfalls lösen, jedoch mit einer höheren Wahrscheinlichkeit (eine Person mit $\theta = 8$ hätte eine Lösungswahrscheinlichkeit von über 80%). Umgekehrt führen geringe Personenfähigkeiten zu geringeren Lösungswahrscheinlichkeiten, so wäre die Lösungswahrscheinlichkeit einer Person mit $\theta = 4$, kleiner als 20%.

Unterschiedliche Itemschwierigkeiten führen im Rasch-Modell zu einer Verschiebung der ICC auf der x-Achse (Abb. 4.2), Form und Steigung bleiben jedoch gleich (Strobl, 2012). Dies liegt in der Modellgleichung (Gl. 4.1) begründet, da für jedes Item nur ein Parameter, nämlich die Itemschwierigkeit (β_j), veränderlich ist (Strobl, 2012).

Die Trennschärfe eines Items (engl.: item discriminability) ist die Steigung des mittleren Bereichs der ICC (Strobl, 2012). Ein Item mit hoher Trennschärfe kann gut zwischen Personen mit unterschiedlichen

Fähigkeiten unterscheiden (diskriminieren) (Döring & Bortz, 2016; Kelava & Moosbrugger, 2020; Koller et al., 2012; Rost, 2004; Strobl, 2012). Die Modellgleichung des Rasch-Modells (Gl. 4.1) lässt nur Items mit gleicher Trennschärfe zu, Items, die dieses Kriterium nicht erfüllen, sind nicht ›Rasch-skalierbar‹ (Strobl, 2012) und müssen aus dem Datensatz ausgeschlossen werden.

4.1.1.1 Eigenschaften des Rasch-Modells

Dieses Kapitel stellt die zentralen Eigenschaften des Rasch-Modells vor. Nach Strobl (2012) begründen diese Eigenschaften, »warum das Rasch-Modell aus theoretischer Sicht so ansprechend und inzwischen auch in der Praxis so etabliert ist: Es erlaubt die objektive Messung von latenten Eigenschaften« (Strobl, 2012, S. 14).

Eine Eigenschaft des Rasch-Modells ist, dass für die beiden Unbekannten (Personenfähigkeit und Itemschwierigkeit) der Modellgleichung suffiziente Statistiken vorliegen (Koller et al., 2012; Strobl, 2012). Suffizienz bedeutet in diesem Zusammenhang, dass alle in der Stichprobe enthaltenen Informationen zu den Unbekannten enthalten sind (Bortz & Schuster, 2010). Dazu müssen nicht die vollständigen Daten bekannt sein, für die Personenfähigkeit θ_i einer Person i enthält bereits die Zeilenrandsumme (Summenscore) alle notwendigen Informationen, für die Itemschwierigkeit β_j eines Items j die Spaltenrandsumme (Koller et al., 2012; Strobl, 2012). Bei den Personenfähigkeiten hängt es also nicht davon ab, welche Items gelöst werden, sondern von der Gesamtzahl der gelösten Items (Kelava & Moosbrugger, 2020; Strobl, 2012).

Eine weitere Eigenschaft des Rasch-Modells ist die lokale stochastische Unabhängigkeit. Im Allgemeinen bedeutet stochastische Unabhängigkeit, dass zwei Zufallsereignisse sich nicht gegenseitig bedingen (Koller et al., 2012). Im Rasch-Modell impliziert dies, dass die Wahrscheinlichkeit, ein Item j zu lösen ($u_{ij} = 1$), nicht von einem anderen Item n abhängen darf (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012). Der Begriff ›lokal‹ bedeutet in diesem Zusammenhang, dass stochastische Unabhängigkeit nur bei gleichen Personenfähigkeiten vorliegen muss (Koller et al., 2012; Strobl, 2012). Wenn also Informationen oder gar die Lösung des Items j zur Lösung des Items n benötigt

werden, wäre die Annahme der lokalen stochastischen Unabhängigkeit verletzt, da die Lösungswahrscheinlichkeit des Items n durch das Item j beeinflusst wird. Die lokale stochastische Unabhängigkeit kann mit der Inter-Itemkorrelation geprüft werden, bei der die Korrelationen zwischen zwei beliebigen Itemvariablen (U_{ij}) untersucht werden (Kelava & Moosbrugger, 2020; Koller et al., 2012). Lokale stochastische Unabhängigkeit, bei konstanter Personenfähigkeit θ, ist bei einer Korrelation von annähernd Null gegeben (Kelava & Moosbrugger, 2020). Strobl (2012) bezieht die lokale stochastische Unabhängigkeit zusätzlich auf die Personen. Die Wahrscheinlichkeit, dass eine Person i ein Item löst, darf also nicht davon abhängen, ob die Person m dasselbe Item lösen kann. Ein Beispiel für eine Verletzung dieser Annahme wäre ein mögliches Abschreiben der beiden Personen während der Testung (Strobl, 2012).

Die spezifische Objektivität ist eine weitere Eigenschaft des Rasch-Modells. Wenn die Personenfähigkeiten zweier Personen nicht von den Items abhängen, anhand derer sie verglichen werden, ist die spezifische Objektivität gegeben (Koller et al., 2012; Strobl, 2012). Genauer gesagt: Wenn die Personen i und m miteinander verglichen werden und Person i eine höhere Personenfähigkeit ($\theta_i > \theta_m$) hat, dann ist die Lösungswahrscheinlichkeit von Person i für ein Item immer höher als die von Person m, unabhängig von der Itemschwierigkeit. Dieser Zusammenhang muss auch auf Itemebene bestehen: Vergleicht man zwei Items mit unterschiedlichen Itemschwierigkeiten ($\beta_j > \beta_n$), so muss das leichtere Item (β_n) bei Personen mit verschiedenen Personenfähigkeiten immer eine höhere Lösungswahrscheinlichkeit aufweisen als das schwierigere Item (β_j) (Strobl, 2012).

Häufig wird statt von spezifischer Objektivität von Stichprobenunabhängigkeit gesprochen (Koller et al., 2012; Strobl, 2012). Allerdings weist Strobl (2012) explizit darauf hin, dass die Rasch-Skalierbarkeit der eigenen Stichprobe nicht unbedingt auf jede beliebige Stichprobe übertragbar ist, da Personen aus einer alternativen Stichprobe die Items u.U. anders interpretieren könnten. In diesem Fall wäre keine Subgruppeninvarianz mehr gegeben und dementsprechend würden Personen, die unterschiedlichen Subgruppen angehören (z.B. aufgrund externer Faktoren wie Geschlecht, Alter, Herkunft, etc.), trotz identischer wahrer

Fähigkeiten unterschiedliche Personenfähigkeiten erzielen (Koller et al., 2012). Dieser Sachverhalt – die Itemschwierigkeit unterscheidet sich zwischen verschiedenen Personen, die unterschiedlichen Subgruppen angehören, obwohl sie die gleichen Personenfähigkeiten besitzen – wird als Differential Item Functioning (DIF) bezeichnet und hat zur Folge, dass die Gültigkeit des Rasch-Modells mit jeder neuen Subgruppe überprüft werden muss (Strobl, 2012).

Die Überprüfung einzelner Items auf DIF kann mit dem Wald-Test erfolgen (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012). Beim Wald-Test »wird die aus den zwei Subpopulationen erhaltene Differenz zweier Itemparameterschätzungen an der Summe der quadrierten Standardfehler der Itemparameterschätzungen relativiert« (Kelava und Moosbrugger, 2020, S. 397). Die so berechnete Teststatistik kann einem Signifikanztest unterzogen werden, da sie einer Standardnormalverteilung folgt (Kelava & Moosbrugger, 2020). Ergibt der Signifikanztest signifikante Ergebnisse, ein Item zeigt über die Subgruppen hinweg unterschiedliche Itemschwierigkeiten, liegt DIF für das Item vor (Kelava & Moosbrugger, 2020). Koller et al. (2012), empfehlen bei Vorliegen auffälliger Items immer nur einzelne Items auszuschließen und anschließend das Rasch-Modell neu zu berechnen. Der Grund dafür ist, dass auffällige Items andere Items beeinflussen können (Koller et al., 2012). So kann es vorkommen, dass bei vorliegen auffälliger Items auch modellkonforme Items als Items mit DIF eingeschätzt werden, auch der umgekehrte Fall ist möglich (Koller et al., 2012). Der beschriebene Wald-Test kann auch als globaler Wald-Test das gesamte Testinstrument auf Rasch-Konformität prüfen (Strobl, 2012).

Im Rasch-Modell liegen Personenfähigkeit und Itemschwierigkeit auf einer gemeinsamen latenten Dimension, erkennbar an der entsprechenden Differenz in der Modellgleichung (Gl. 4.1) (Strobl, 2012). Diese Eindimensionalität bedeutet, dass die verwendeten Items nur ein latentes Merkmal messen (Koller et al., 2012; Strobl, 2012). Items werden als homogen bezeichnet, wenn sie dasselbe latente Merkmal erfassen (Koller et al., 2012). Ist dies der Fall, korrelieren diese Items positiv miteinander (Koller et al., 2012). Koller et al. (2012) weisen jedoch daraufhin, dass zu hohe Korrelationen eine Verletzung der lokalen stochastischen

Unabhängigkeit bedeuten, während heterogene Items zu niedrige Korrelationswerte aufweisen (Koller et al., 2012). Eine klare Abgrenzung zum oben genannten DIF, ist nach Strobl (2012) bei Verletzung der Eindimensionalität nicht eindeutig möglich.

4.1.1.2 Parameterschätzung

In der Praxis wird die Modellgleichung (Gl. 4.1) nicht zur Bestimmung der Lösungswahrscheinlichkeit verwendet, sondern die Personenfähigkeit und die Itemschwierigkeit werden aus der Antwort einer Person auf ein konkretes Item abgeleitet (Koller et al., 2012). Aufgrund der Anzahl der Unbekannten in der Modellgleichung (Gl. 4.1) ist es jedoch nicht möglich, diese Gleichung eindeutig zu lösen, weshalb iterative Schätzverfahren eingesetzt werden, um die Personenfähigkeit und die Itemschwierigkeit entsprechend zu schätzen (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012).

Die meisten Schätzverfahren basieren auf der *Maximum-Likelihood-Methode* (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012). Ausgangspunkt dieser Methode ist die Likelihood-Funktion, welche »die Wahrscheinlichkeit der Daten in Abhängigkeit von den Modellparametern unter der Annahme, dass das Modell gilt« (Rost, 2004, S. 112), beschreibt. Wird die Modellgleichung (Gl. 4.1) als Funktion der unbekannten Parameter dargestellt, so erhält man die Likelihood-Funktion (Strobl, 2012):

$$L_{u_i}(\theta_i, \beta) = \prod_{j=1}^{m} \frac{e^{u_{ij}(\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}} \tag{4.3}$$

Mit dieser Gleichung kann nun aus den vorliegenden Daten (u_{ij}) die Wahrscheinlichkeit/Likelihood (L_{ui}) für die Personenfähigkeit (θ) bzw. die Itemschwierigkeit (β) berechnet werden (Koller et al., 2012). Bei der Parameterschätzung werden Werte für θ und β verwendet, sodass die Likelihood L_{ui} möglichst groß (maximal) wird und damit am besten zu den Daten passt (Koller et al., 2012). Dieses Verfahren wird als *Maximum-Likelihood-Methode* bezeichnet (Koller et al., 2012), eine ausführliche Beschreibung findet sich in Rose (2020, Kap. 19.2), Rost (2004, Kap. 4.1) und Strobl (2012, Kap. B.2).

Die Schätzverfahren unterscheiden sich darin, ob Personenfähig-

keit und Itemschwierigkeit gemeinsam (wie bei der *Joint Maximum Likelihood*) oder nacheinander (*Conditional Maximum Likelihood* oder *Marginal Maximum Likelihood*) geschätzt werden (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012).

Bei der *Joint Maximum Likelihood*-Schätzung (JML) werden die Personenfähigkeit und die Itemschwierigkeit gleichzeitig aus der gemeinsamen Likelihood geschätzt (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012). Koller et al. (2012) und Strobl (2012) raten jedoch aufgrund des *Incidential Parameter Problems* von der Anwendung ab. Generell sind möglichst große Stichproben bei statistischen Betrachtungen wünschenswert, da sie die Varianz der Schätzung reduzierten, dies wird als Konsistenz bezeichnet (Strobl, 2012). Bei der gleichzeitigen Schätzung von Personenfähigkeit und Itemschwierigkeit ist dies jedoch nicht der Fall, da in der Regel die Anzahl der Items durch das Testdesign festgelegt ist und nur die Anzahl der Personen erhöht werden kann. Eine höhere Personenzahl erhöht allerdings die Anzahl der unbekannten Personenfähigkeiten, die aufgrund der gleichbleibenden Itemzahl nicht genauer geschätzt werden können (Koller et al., 2012). Die gemeinsame Schätzung führt zu fehlerbehafteten Itemschwierigkeiten (Koller et al., 2012), weshalb die Anwendung eines zweistufigen Schätzverfahrens zur getrennten Schätzung von Personenfähigkeit und Itemschwierigkeit empfohlen wird (Koller et al., 2012; Strobl, 2012).

Bei den Verfahren *Conditional Maximum Likelihood* (CML) und *Marginal Maximum Likelihood* (MML) werden zunächst die Itemschwierigkeiten geschätzt, die Schätzung der Personenfähigkeiten erfolgt erst in einem zweiten Schritt (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012). Das *Conditional Maximum Likelihood*-Verfahren nutzt die im Rasch-Modell als suffiziente Statistik geltenden Randsummen der Personen (Summenscores), um die Personenparameter aus der Likelihood-Funktion (Gl. 4.3) zu kürzen (Kelava & Moosbrugger, 2020; Koller et al., 2012; Rost, 2004; Strobl, 2012). Dies ist auch der Grund, warum dieses Verfahren nur im Rasch-Modell zulässig ist (Rost, 2004). Ein weiterer Nachteil besteht darin, dass die zuvor geschätzten Itemschwierigkeiten bei der Schätzung der Personenfähigkeiten als wahre Werte angenommen werden, dadurch werden die durch die Schätzung

verursachte Unsicherheiten ignoriert (Strobl, 2012). Ebenso werden bei diesem Schätzverfahren keine Personenfähigkeiten für Personen geschätzt, die keine oder alle Items richtig gelöst haben (Strobl, 2012).

Anstelle von Randsummen wird bei dem *Marginal Maximum Likelihood*-Verfahren eine Randverteilung für die Personenfähigkeiten angenommen (Kelava & Moosbrugger, 2020; Strobl, 2012). In der Regel handelt es sich um eine Standard-Normalverteilung mit einem Erwartungswert von 0 und einer Varianz von 1 (Kelava & Moosbrugger, 2020; Strobl, 2012). Das Verfahren kann sowohl für das Rasch-Modell als auch für mehrparametrische Modelle (2pl- oder 3pl-Modell) verwendet werden (Kelava & Moosbrugger, 2020). Bei der Schätzung der Personenfähigkeiten hat auch dieses Verfahren den Nachteil, dass die zuvor geschätzten Itemschwierigkeiten als wahre Werte angenommen werden, ohne diese Unsicherheiten zu berücksichtigen (Strobl, 2012). Auch kann die Normalverteilung als Randverteilung unpassend sein, »wenn eine Personen-Stichprobe nicht aus der Normalbevölkerung stammt« (Strobl, 2012, S. 34). Vorteilhaft ist hingegen, dass die zu schätzenden Parameter nicht vom Stichprobenumfang abhängen (Rose, 2020). Nach Rose (2020) ist das *Marginal Maximum Likelihood*-Verfahren »die am häufigsten verwendete Schätzmethode in der IRT« (Rose, 2020, S. 449).

Weitere Schätzansätze, die nicht auf der *Maximum-Likelihood*-Methode basieren, wie *Bayesianische*-Ansätze, die auf die *Markov-Chain-Monte-Carlo*-Methode zurückgreifen, werden in dieser Arbeit nicht verwendet, eine ausführliche Darstellung findet sich in Rose (2020, Kap. 19.3).

Für die Schätzung der Personenfähigkeiten stehen verschiedene Verfahren zur Verfügung. Wie bereits erwähnt, werden bei der *Joint Maximum Likelihood*-Schätzung die Personenfähigkeit und die Itemschwierigkeit gleichzeitig geschätzt (Kelava & Moosbrugger, 2020; Koller et al., 2012; Strobl, 2012). Nach Rost (2004) erweist sich dieser Schätzer allerdings nur dann als konsistent, »wenn die Anzahl der Items gegen unendlich geht, d.h. relativ groß ist« (Rost, 2004, S. 312).

Wurde die Itemschwierigkeit mit der *Conditional Maximum Likelihood*- oder der *Marginal Maximum Likelihood*-Methode geschätzt, können die Personenfähigkeiten mit dem *Maximum Likelihood Estima-*

tes-Schätzer (MLE) oder dem *Weighted Maximum Likelihood Estimates*-Schätzer (WLE) geschätzt werden (Kelava & Moosbrugger, 2020; Rost, 2004). Bei den *Maximum Likelihood Estimates* werden die Personenfähigkeiten, wie oben bereits erläutert, über die im Rasch-Modell als suffiziente Statistik geltenden Randsummen der Personen (Summenscores) geschätzt (Koller et al., 2012; Strobl, 2012). Eine Schätzung von Personen mit Extremwerten (kein Item oder alle Items beantwortet) ist nicht möglich (Kelava & Moosbrugger, 2020; Koller et al., 2012; Rost, 2004; Strobl, 2012). Abhilfe kann das Verfahren der *Weighted Maximum Likelihood Estimates* schaffen. Nach dem Bayes'schen Ansatz der Parameterschätzung wird die Wahrscheinlichkeit der Personenfähigkeit $P(\theta_i \mid u_{ij}, \beta_j)$ maximiert und nicht wie bei den *Maximum Likelihood*-Verfahren die Wahrscheinlichkeit der Daten $P(u_{ij} \mid \theta_i, \beta_j)$ (Rost, 2004). Im Ergebnis ist die Varianz der *Weighted Maximum Likelihood Estimates* geringer als die der *Maximum Likelihood Estimates* (Rost, 2004) und »hat diese als ›Standardverfahren‹ ersetzt« (Rost, 2004, S. 314). Die *Weighted Maximum Likelihood Estimates* können auch die Personenfähigkeiten von Personen mit Extremwerten schätzen, insgesamt liefern sie »weniger verzerrte Personenparameterschätzungen für endliche Stichproben« (Kelava und Moosbrugger, 2020, S. 393).

Ein weiteres Verfahren stellt der *Expected a Posteriori*-Schätzer (EAP) dar, der keine individuellen Schätzwerte, sondern Erwartungswerte der Personenfähigkeiten aus den im Rahmen der *Marginal Maximum Likelihood* geschätzten Verteilungsparametern der latenten Variablen bestimmt (Rost, 2004). Die Personenfähigkeiten sind Mittelwerte einer kleinen Verteilung, aus der bei der *Plausible Values*-Methode Zufallswerte gezogen werden (Rost, 2004). *Plausible Values* werden in Large-Scale-Assessments wie PISA verwendet (siehe Kap. 2.1.2).

Im Anschluss einer ersten Rasch-Skalierung müssen die Items hinsichtlich Gütekriterien und ihrer Modellpassung untersucht werden.

4.1.2 Prüfung der Item- und Testgüte

Die in einer Rasch-Skalierung verwendeten Items müssen den Anforderungen des Rasch-Modells genügen. Daher ist eine Prüfung des Item-Fits und der Reliabilität notwendig. Diese beiden Kriterien werden in diesem

Kapitel beschrieben.

Der Item-Fit gibt an, wie gut ein Item in das Modell passt (Wilson, 2023). Dabei wird anhand der Residuen überprüft, ob sich die beobachteten und die vom Modell vorhergesagten Lösungswahrscheinlichkeiten unterscheiden (K. Müller et al., 2017). Durch die Überprüfung der Fit-Werte kann verhindert werden, dass unpassende Items die Schätzungen verzerren (Wilson, 2023). In der Regel werden zwei Item-Fit-Werte bestimmt: Infit (*Weighted-Mean-Square* [wMNSQ-Infit]) und Outfit (*Unweighted-Mean-Square* [wMNSQ-Outfit]) (Bond et al., 2020; J. Müller et al., 2018; Wilson, 2023). Der wMNSQ-Outfit wird nach der gleichen Berechnung wie der wMNSQ-Infit bestimmt, jedoch werden die Personen, deren Personenfähigkeiten eine größere Differenz zur Itemschwierigkeit aufweisen (Randbereiche), stärker berücksichtigt (Wilson, 2023). Wenn das Item in das Modell passt, dann sollte der wMNSQ = 1 sein (Wilson, 2023). Grenzwerte für den wMNSQ variieren in der Literatur, Wilson (2023) nennt Werte zwischen 0.75 und 1.33 als angemessen, während in PISA Werte zwischen 0.8 und 1.2 akzeptabel sind (K. Müller et al., 2017). In dieser Arbeit werden die Grenzen der PISA-Studie (K. Müller et al., 2017) verwendet. Wenn ein wMNSQ-Wert außerhalb dieser Grenzen liegt, wird ein weiterer Fit-Wert mit dem gewichteten t betrachtet. Das gewichtete t verwendet eine Transformation, um den wMNSQ in eine Standard-Normalverteilung umzuwandeln, um die Signifikanz zu prüfen (Wilson, 2023). Liegt t außerhalb der Grenzen von -2 und 2, sollte das Item ausgeschlossen werden (Bond et al., 2020).

Das letzte in diesem Kapitel vorgestellte Gütekriterium ist die Reliabilität. In der KTT wird die Reliabilität der manifesten Testvariablen als Varianzverhältnis zwischen der Varianz des wahren Testwertes und der Varianz der mainfesten Testvariablen definiert, der Standardmessfehler dient ebenfalls zur Beschreibung der Messgenauigkeit (Rose, 2020). Auch in der IRT kann die Reliabilität zur Beschreibung der Messgenauigkeit herangezogen werden, die Berechnung erfolgt jedoch abweichend durch Zerlegung des Personenfähigkeitsschätzers (Rose, 2020). In der IRT ist es im Gegensatz zur KTT schwierig einen Kennwert für die Reliabilität anzugeben, »da die Standardfehler der Personenparameter-

schätzungen in Abhängigkeit der latenten Personenvariablen variieren« (Rose, 2020, S. 494). Marginale Reliabilitätsmaße können als mittlere Reliabilität (über die latente Variable) interpretiert werden (Rose, 2020). Die Berechnung der Reliabilitäten und Standardfehler unterscheidet sich je nach verwendetem Personenfähigkeitsschätzer, eine Übersicht gibt Rose (2020). Die Reliabilitäten können sowohl für Personenfähigkeiten (*Person Separation Reliability*), als auch für Itemschwierigkeiten (*Item Separation Reliability*) berechnet werden und lassen sich wie Cronbach´s α interpretieren (Bond et al., 2020). Zur Interpretation von Cronbach´s α gibt Blanz (2015) folgende Abstufungen an:

- Cronbach´s $\alpha < .5$ - inakzeptabel
- Cronbach´s $\alpha > .5$ - schlecht (niedrig)
- Cronbach´s $\alpha > .6$ - fragwürdig
- Cronbach´s $\alpha > .7$ - akzeptabel
- Cronbach´s $\alpha > .8$ - gut (hoch)
- Cronbach´s $\alpha > .9$ - exzellent (Blanz, 2015).

Inhaltlich bedeuten niedrige Reliabilitäten, dass die Personenfähigkeit bzw. die Itemschwierigkeit nicht zuverlässig bestimmt werden kann und dementsprechend zu vermeiden sind.

Nach der Überprüfung der Item- und Testgüte kann untersucht werden, ob die erhobenen Daten dem Rasch-Modell (1pl) entsprechen. Alternative Modelle werden im folgenden Kapitel vorgestellt.

4.1.3 2pl/3pl-Modell nach Birnbaum und mehrdimensionale Modelle

Neben dem in Kap. 4.1.1 beschriebenen Rasch-Modell (1pl) gibt es zahlreiche Erweiterungen. Zwei dieser Erweiterungen, die Birnbaum-Modelle 2pl und 3pl, werden im Folgenden erläutert.

Neben den dichotomen 1pl-, 2pl- und 3pl-Modellen existieren in der IRT weitere Modelle, die sich »u.a. durch die Art der manifesten und/

oder latenten Variablen und die Art der verwendeten Modellparameter« (Kelava und Moosbrugger, 2020, S. 403) unterscheiden und bei denen lokale stochastische Unabhängigkeit vorliegt (Kelava et al., 2020). Auf eine Darstellung dieser Modelle, die bei mehrstufigen Antwortkategorien (u.a. *Partial-Credit-Modell* oder *Raing-Scale-Modell*) oder zur Modellierung von Unterschieden zwischen Personen (u.a. *Mixed-Rasch-Modell*) eingesetzt werden können, wird verzichtet. Erläuterungen hierzu sind zu finden in Kelava et al. (2020), Kelava und Moosbrugger (2020), Rost (2004, Kap. 3) und Strobl (2012, Kap. 5).

Im 1pl-Modell nach Rasch (siehe Kap. 4.1) gilt die Bedingung, dass alle Items die gleiche Trennschärfe besitzen, mit der Folge, dass die Steigung der ICC gleich ist (Kelava & Moosbrugger, 2020; Rost, 2004; Strobl, 2012). Diese Bedingung wird in den beiden Birnbaum-Modellen gelockert, sodass unterschiedliche Trennschärfen der Items möglich sind und somit auch die Steigungen der ICC variieren können (Kelava & Moosbrugger, 2020; Rost, 2004; Strobl, 2012). Neben den beiden bekannten Parametern Personenfähigkeit θ_i und Itemschwierigkeit β_j wird mit dem Diskriminationsparameter δ_j (auch Trennschärfeparameter oder Steigungsparameter genannt) ein weiterer Itemparameter eingeführt, der die Steigung der ICC beeinflusst (Kelava & Moosbrugger, 2020; Rost, 2004; Strobl, 2012):

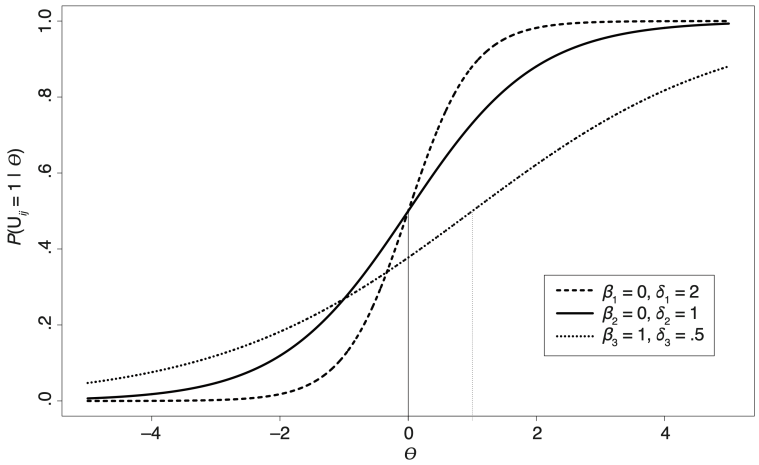
$$P\left(u_{ij} = 1 \mid \theta_i, \beta_j, \delta_j\right) = \frac{e^{\delta_j(\theta_i - \beta_j)}}{1 + e^{\delta_j(\theta_i - \beta_j)}} \tag{4.4}$$

Der Diskriminationsparameter δ_j in Gl. 4.4 führt dazu, dass man »unterschiedlich gut zwischen Personen mit schwächeren bzw. stärkeren Merkmalsausprägungen trennen« (Kelava und Moosbrugger, 2020, S. 399) kann, sodass sich die ICC von Items überschneiden können und somit keine spezifische Objektivität mehr vorliegt (Strobl, 2012).

Abb. 4.3 zeigt die ICC für drei Items. Die Itemschwierigkeit der Items 1 & 2 beträgt jeweils $\beta_{1;2} = 0$, jedoch unterscheiden sich die Diskriminationsparameter ($\delta_1 = 2$ bzw. $\delta_2 = 1$).

Für Personen mit einer Personenfähigkeit von $\theta = 0$ liegt die Lösungswahrscheinlichkeit der Items mit einer Itemschwierigkeit von $\beta_{1;2} = 0$ bei 50%, wie aus dem Rasch-Modell bekannt, da die Itemschwierigkeiten

Abbildung 4.3
Itemchararakteristikkurve für mehrere Items mit unterschiedlichen Diskriminationsparametern (Kelava & Moosbrugger, 2020, S. 400, Bezeichnungen angepasst)



Anmerkungen. $P(U_{ij} = 1 \mid \theta)$ = Lösungswahrscheinlichkeit; θ = Personenfähigkeit; β = Itemschwierigkeit; δ = Diskriminationsparameter.

identisch sind. Betrachtet man nun allerdings eine Person mit einer geringeren Personenfähigkeit von $\theta = -2$, so wird der Einfluss des Diskiriminationsparameter deutlich, da die Lösungswahrscheinlichkeit von Item 1 praktisch bei 0% liegt, während Item 2 mit einer Wahrscheinlichkeit von ca. 10% gelöst wird. Bei Betrachtung des dritten Items wird der Einfluss nochmals deutlich. Die Itemschwierigkeit liegt mit $\beta_3 = 1$ über den beiden anderen Items, da aber der Diskiriminationsparameter mit $\delta_3 = .5$ unter 1 liegt (flacher Verlauf der ICC), ist die Lösungswahrscheinlichkeit bei einer Personenfähigkeit von $\theta < -1$ größer als bei den beiden anderen Items mit kleineren Itemschwierigkeiten. Somit können im Birnbaum-Modell (2pl) »Items unterschiedliche Reihenfolgen ihrer Lösungswahrscheinlichkeiten aufweisen« (Rost, 2004, S. 134).

Da die Randsummen im Birnbaum-Modell (2pl) keine suffiziente Statistik darstellen, muss zur Schätzung der Itemschwierigkeit das *Marginal Maximum Likelihood-Verfahren* (siehe Kap. 4.1.1.2) verwendet werden (Strobl, 2012).

Mit dem 3pl-Modell nach Birnbaum wird ein weiterer Itemparameter – der Rateparameter γ_j – eingeführt, weshalb das 3pl-Modell auch als

Ratemodell bezeichnet wird (Kelava & Moosbrugger, 2020; Strobl, 2012). Der Rateparameter γ_j bestimmt, ob die ICC einer Aufgabe wie üblich bei 0 beginnt oder einen größeren Wert annimmt (Strobl, 2012). Ist der Rateparameter $\gamma_j > 0$, so kann auch eine Person mit einer unendlich kleinen Personenfähigkeit dieses Item mit einer Wahrscheinlichkeit $P > 0$ durch Raten lösen, bei Multiple-Choice-Aufgaben wäre dies der Fall (Kelava & Moosbrugger, 2020; Strobl, 2012). Die Gleichung des 3pl-Modells nach Birnbaum lautet (Strobl, 2012):

$$P\left(u_{ij} = 1 \mid \theta_i, \beta_j, \delta_j, \gamma_j\right) = \gamma_j + (1 - \gamma_j) \cdot \left(\frac{e^{\delta_j(\theta_i - \beta_j)}}{1 + e^{\delta_j(\theta_i - \beta_j)}}\right) \tag{4.5}$$

Da ein weiterer Parameter (Unbekannte) in die Gleichung (Gl. 4.5) eingeführt wird, ist eine sehr große Stichprobe erforderlich (Strobl, 2012). Das 2pl-Modell nach Birnbaum ist ein Spezialfall des 3pl-Birnbaum-Modells, nämlich wenn der Rateparameter $\gamma_j = 0$ ist (Strobl, 2012).

Kelava und Moosbrugger (2020) weisen darauf hin, dass zwar mit dem 2pl- und 3pl-Birnbaum-Modell genauere Modellierungen möglich sind, aber nur das Rasch-Modell (1pl) die Vorteile der suffizienten Statistik, der spezifischen Objektivität und der Schätzverfahren bietet und aufgrund der geringeren Parameterzahl auch ökonomischer/sparsamer ist.

Neben den bisher diskutierten Erweiterungen des Rasch-Modells, die der Grundannahme der Eindimensionalität folgen, existieren auch mehrdimensionale IRT-Modelle (kurz MIRT-Modelle) (Kelava et al., 2020; Rost, 2004; Strobl, 2012). Generell modelliert ein mehrdimensionales Modell das vorliegende Kompetenzkonstrukt auf der Basis mehrerer Teilkompetenzen (Hartig & Höhler, 2010). Im Vergleich zur eindimensionalen Modellierung ermöglicht das mehrdimensionale Modell »eine differenziertere Diagnostik und zugleich eine Prüfung von Annahmen über die Struktur der erfassten Kompetenz und Teilkompetenzen« (Hartig und Höhler, 2010, S. 190). Bei mehrdimensionalen Modellen wird unterschieden, ob ein Item einer oder mehreren Dimensionen angehört (Strobl, 2012). Lädt ein Item nur auf einer Dimension, z.B. erfasst das Item beispielsweise nur die Rechenfähigkeit, so spricht man von einer *Einfachstruktur* oder von *Within-Item-Multidimensionality* (Har-

tig & Höhler, 2010; Strobl, 2012). Alternativ können Items auch auf mehreren Dimensionen laden, d.h. inhaltlich zwei oder mehr Kompetenzen gleichzeitig abdecken, in diesem Fall spricht man von *Within-Item Multidimensionality* (Hartig & Höhler, 2010; Strobl, 2012).

Welches der vorgestellten Modelle am besten zu den erhobenen Daten passt, muss durch einen Modellvergleich ermittelt werden.

4.1.4 Modellvergleiche

Modellvergleiche dienen dazu, die Passung der erhobenen Daten zwischen verschiedenen Modellen zu überprüfen. Meist sollen diese Modellvergleichstests zeigen, dass die geschätzten Itemschwierigkeiten keine systematischen Unterschiede zwischen den Personengruppen aufweisen (Strobl, 2012). Nach Strobl (2012) testen diese Modellvergleichstests also in der Regel auf DIF, das wiederum auf eine Mehrdimensionalität des Tests hinweisen kann. Neben einem graphischen Modelltest – der in dieser Arbeit nicht verwendet wird – existieren verschiedene statistische Tests zur Modellprüfung, wie der *Likelihood-Ratio-Test*, der χ^2-*Anpassungstest* mit *Bootstrap-Verfahren* oder Informationskriterien (Gollwitzer, 2020; Strobl, 2012).

Der *Likelihood-Ratio-Test* (LRT), auch *Likelihood-Quotienten-Test* genannt, beruht auf der Annahme, dass sich in einem Rasch-Modell die Itemschwierigkeiten in unterschiedlichen Personengruppen nicht unterscheiden dürfen (Strobl, 2012). Dies ist auch die Grundlage des graphischen Modelltests, bei dem allerdings nur zwei Gruppen direkt miteinander verglichen werden, während bei dem *Likelihood-Quotienten-Test* beliebig viele Gruppen möglich sind (Strobl, 2012). Dazu werden die Personen in Gruppen anhand ihrer Rohwerte oder anderer Merkmale (wie z.B. Geschlecht) eingeteilt, anschließend werden die Itemschwierigkeiten für alle Gruppen geschätzt (Strobl, 2012). Die geschätzten Itemschwierigkeiten werden in die jeweilige Likelihood-Funktion eingesetzt, die abschließend miteinander verglichen werden (Likelihood-Quotient) (Gollwitzer, 2020; Strobl, 2012). Die Nullhypothese des *Likelihood-Quotienten-Test* besagt, dass alle Personen die gleichen Aufgabenschwierigkeiten haben, das Rasch-Modell also erfüllt ist (Strobl, 2012). Ist der Likelihood-Quotient kleiner als 1, so gilt die Alternativhypothese,

dass für Personengruppen unterschiedliche Itemschwierigkeiten gelten und somit die Bedingung des Rasch-Modells nicht erfüllt ist (Strobl, 2012). Das Ergebnis muss noch auf Signifikanz geprüft werden, indem der Likelihood-Quotient in eine Prüfgröße transformiert wird, die einer χ^2-Verteilung folgt (Gollwitzer, 2020; Strobl, 2012).

Bei dem χ^2-*Anpassungstest* werden die Häufigkeiten der geschätzten Antwortmuster mit den beobachteten Häufigkeiten verglichen (Gollwitzer, 2020; Strobl, 2012). Dieses Verfahren kann jedoch nur angewendet werden, wenn die Häufigkeit der beobachteten Antwortmuster nicht zu klein ist (Gollwitzer, 2020; Strobl, 2012). Dies kann der Fall sein, wenn die Anzahl der Items zu hoch ist, da dann bestimmte Antwortmuster nur selten auftreten (Strobl, 2012), aber auch wenn die Stichprobengröße zu klein ist (Gollwitzer, 2020). In diesen Fällen kann mit dem *Bootstrap-Verfahren* eine simulierte Prüfverteilung mit einer großen Anzahl künstlicher Datensätze generiert werden (Gollwitzer, 2020). Aus diesen lässt sich wiederum die Teststatistik berechnen, die zur Annahme/Ablehnung des Modells führen kann (Gollwitzer, 2020; Strobl, 2012).

Akaike Information Criterion (AIC), *Bayesian Information Criterion* (BIC) und *Consistent Akaike Information Criterion* (CAIC) können als Informationskriterien für den Modellvergleich verwendet werden (Gollwitzer, 2020). Diese Informationsparameter basieren ebenfalls auf der Likelihood-Funktion, sie berücksichtigen die Anzahl der Modellparameter und ›bestrafen‹ zu viele oder unnötige Parameter (Gollwitzer, 2020). Vereinfacht ausgedrückt werden sparsame Modelle belohnt (niedriger Wert des Informationskriteriums), während komplexe Modelle bestraft werden (hoher Wert des Informationskriteriums) (Gollwitzer, 2020). Dementsprechend gilt für alle drei Informationskriterien, dass das Modell mit dem niedrigeren Wert am besten passt (Rost, 2004). Alle Informationskriterien beinhalten die (logarithmierte) Likelihood $LogLike. = ln(L)$, die auch als $Deviance = -2 \cdot \ln(L)$ angegeben wird, und die Anzahl der Modellparameter t:

$$AIC = -2 \cdot \ln(L) + 2 \cdot t \tag{4.6}$$

$$BIC = -2 \cdot \ln(L) + t \cdot \ln(n) \tag{4.7}$$

$$CAIC = -2 \cdot \ln(L) + t \cdot (\ln(n) + 1) \tag{4.8}$$

Bei BIC (Gl. 4.7) und CAIC (Gl. 4.8) wird zusätzlich die Stichprobengröße n berücksichtigt. Rost (2004) empfiehlt die Verwendung des AIC (Gl. 4.6) bei kleinen Stichproben, des BIC (Gl. 4.7) bei großen Stichproben und des CAIC (Gl. 4.8) bei sehr großen Stichproben, ohne die jeweiligen Stichprobengrößen näher zu quantifizieren.

Die IRT skalierten Daten können zur Modellierung von Kompetenzniveaus verwendet werden. Das entsprechende Verfahren wird im nächsten Kapitel beschrieben.

4.2 Niveaumodellierung

Die Antworten der Probanden in den Testinstrumenten, werden durch eine Fähigkeit hervorgerufen (Rauch & Hartig, 2020). Diese Fähigkeit (latente Variable) kann nicht direkt gemessen werden, sondern wird aus dem Antwortverhalten der Probanden (manifeste Variable) abgeleitet (Rauch & Hartig, 2020). Dies geschieht durch die in Kap. 4.1 beschriebene Anwendung von IRT-Modellen auf die vorliegenden Daten. Die so gewonnenen Personenfähigkeiten ermöglichen einen »normorientierte[n] Vergleich von Leistungswerten mit Bezugspopulationen oder der Vergleich von Subpopulationen untereinander« (Rauch und Hartig, 2020, S. 413).

In der (schulischen) Kompetenzdiagnostik (siehe Kap. 2.1.2) sind solche Vergleiche nicht mehr ausreichend und zeitgemäß, weshalb kriteriumsorientierte Interpretationen von Testwerten in den Fokus rücken (Rauch & Hartig, 2020). Die kriteriumsorientierte Interpretation von Testwerten ermöglicht es, bestimmte Probandengruppen zu identifizieren, die über spezifische Kompetenzen verfügen, um fachbezogene Leistungsanforderungen mit hinreichender Sicherheit zu erfüllen (Rauch & Hartig, 2020). Dazu müssen konkrete fachliche Anforderungen mit den Testwerten in Bezug gesetzt werden (Rauch & Hartig, 2020). Die

konkrete fachliche Beschreibung jedes Punktes auf der kontinuierlichen Kompetenzskala ist jedoch praktisch nicht umsetzbar (Beaton & Allen, 1992). Daher werden die empirisch erfassten Kompetenzen durch Kompetenzniveaumodelle inhaltlich konkret beschrieben (Hartig & Klieme, 2006). Häufig wird in der Bildungsforschung aus pragmatischen Gründen die kontinuierliche Kompetenzskala in Abschnitte unterteilt, die als Kompetenzniveaus oder Kompetenzstufen bezeichnet und jeweils kriteriumsorientiert beschrieben werden (Hartig & Klieme, 2006). Kriteriumsorientiert meint in diesem Zusammenhang die schwierigkeitsbestimmenden Merkmale der Items (Dammann, 2016). In der Literatur wird häufig der Begriff ›Kompetenzstufen‹ verwendet, um die einzelnen Abschnitte auf der kontinuierlichen Skala zu beschreiben. Hartig und Klieme (2006) sehen die Verwendung dieser Begrirfflichkeit kritisch und empfehlen daher die Verwendung des Begriffs ›Kompetenzniveau‹, der auch in dieser Arbeit verwendet wird. Von Rauch und Hartig (2020) werden die normorientierte und die kriteriumsorientierte Interpretation von Leistungswerten nicht in Konkurrenz zueinander verstanden, sondern als Alternative oder Ergänzung.

Da Kompetenzniveaumodelle nicht nur im schulischen Kontext von Interesse sind, werden im Rahmen dieser Arbeit entsprechende Niveaumodelle für die Technischen Mechanik und die Rechenfähigkeit für den Studiengang Bauingenieurwesen konzipiert.

4.2.1 Grundlagen der Niveaumodellierung

Die Skalierung auf einer gemeinsamen Skala von Itemschwierigkeiten und Personenfähigkeiten ist nach Rauch und Hartig (2020) die Grundvoraussetzung für eine kriteriumsorientierte Interpretation individueller Testwerte. Diese Grundvoraussetzung ist durch die IRT-Skalierung (siehe Kap. 4.1) gegeben, die eine Interpretation der individuellen Personenfähigkeiten durch ihre Abstände zu den Itemschwierigkeiten ermöglicht (Rauch & Hartig, 2020). Dies ist allerdings nur dann gültig, wenn die ICC aller Items parallel verlaufen, also die spezifische Objektivität (siehe Kap. 4.1.1.1) gegeben ist (Rauch & Hartig, 2020). Die spezifische Objektivität, d.h. die Beschreibung der ICC-Funktion durch einen einzelnen Parameter (Itemschwierigkeit) ist allerdings nur beim 1pl-Rasch-Modell

gegeben (siehe Kap. 4.1.1.1). Damit würde das 2pl- oder 3pl-Birnbaum-Modell für die Niveaumodellierung ausscheiden. Wilson (2003) nennt jedoch vier mögliche Lösungsansätze, um diesem Problem zu begegnen:

1. Entwicklung einer komplexeren Interpretation des Konstrukts:
   Diese komplexe Interpretation könnte die wechselnden Rangfolgen der Itemschwierigkeiten einbeziehen und erklären. Die Tatsache, dass der Abstand zwischen zwei Personen für alle Items gleich ist, könnte im 2pl-Birnbaum-Modell dadurch ersetzt werden, dass das Verhältnis dieser Abstände für jedes Itempaar konstant ist. Es ist jedoch unklar, wie diese Eigenschaft sinnvoll interpretiert werden kann.

2. Verwendung der ›Response Probability 80‹:
   Hier wird eine Lösungswahrscheinlichkeit von 80% angenommen. Die inkonsistente Rangfolge der Itemschwierigkeiten, die sich bei geringeren Lösungswahrscheinlichkeiten ergibt (siehe Kap. 4.1.3), wird bei diesem Ansatz ignoriert.

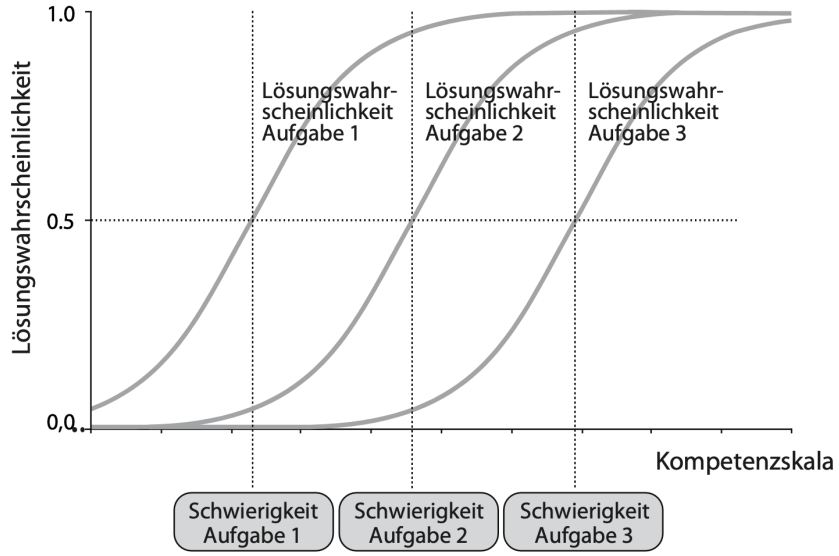
3. Ignorieren der inhaltlichen Interpretierbarkeit der Itemschwierigkeiten:
   Im Hinblick auf die interne Konstruktvalidität wird auf eine Interpretation der Itemschwierigkeiten und Lösungswahrscheinlichkeiten verzichtet, die Daten werden so akzeptiert, wie sie sind. Dies widerspricht jedoch den aktuellen Teststandards.

4. Anzahl der Items erhöhen:
   Es können mehr Items entwickelt werden, wenn die Items das Konstrukt besser abbilden, sodass sich die ICC nicht überlagern und entsprechend das 1pl-Rasch-Modell gilt (Wilson, 2003).

Bei Lösungsansatz 1 ist unklar, ob eine sinnvolle und nützliche Interpretation des komplexen Konstrukts möglich ist. Der dritte Ansatz entspricht nicht den aktuellen Teststandards und sollte daher nicht verwendet werden. Lösungsansatz 4 ist nur umsetzbar, wenn die Datenerhebung bzw. weitere Erhebungen noch ausstehen. Somit erscheint nur Lösungsansatz 3 praktikabel, sofern kein 1pl-Rasch-Modell vorliegt.

Bereits in Kap. 4.1.1 wurde erläutert, dass im 1pl-Rasch-Modell die Personenfähigkeit gleich der Itemschwierigkeit an der Stelle ist, an der ein Item mit einer Wahrscheinlichkeit von 50% gelöst wird. Dies zeigt sich auch in den ICC (Abb. 4.4), wo der Wendepunkt genau diese Itemschwierigkeit darstellt.

Abbildung 4.4
Verankerung von Testaufgaben auf der Kompetenzskala (Hartig & Klieme, 2006, S. 134)

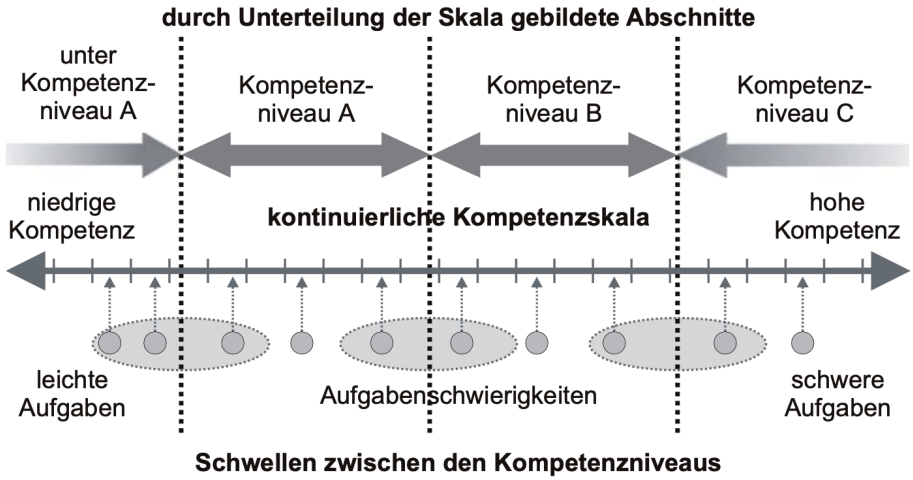


Probanden, deren Personenfähigkeit diese Schwelle von 50% eines Items (Itemschwierigkeit) überschreitet, können dieses Item mit hinreichender Wahrscheinlichkeit ($P(u_{ij} = 1) \geq 50\%$) richtig lösen (Rauch & Hartig, 2020). Eine kriteriumsorientierte Beschreibung der Anforderungen der Items, die ein Proband mit einer Wahrscheinlichkeit von 50% lösen kann, kann auf der Basis der Anforderungen erfolgen, die in den Items zur Beantwortung benötigt werden (Rauch & Hartig, 2020). Allerdings stellen Rauch und Hartig (2020) diese Lösungswahrscheinlichkeit von 50% in Frage, da sie als relativ niedrig angesehen wird, damit die Probanden die Anforderungen mit hinreichender Sicherheit bewältigen können. Es wird daher empfohlen, eine höhere Lösungswahrscheinlichkeit von z.B. 65% anzusetzen (Rauch & Hartig, 2020), wie

sie auch in anderen Studien verwendet wird (vgl. Baumert et al., 1999; Hartig, 2007). Beaton und Allen (1992) nennen dagegen eine Lösungswahrscheinlichkeit von 80% als hinreichend, die auch Wilson (2003) für mehrparametrische Modelle nennt.

Um eine kompetenzorientierte Beschreibung der Kompetenzskala zu ermöglichen, wird die kontinuierliche Kompetenzskala aus pragmatischen Gründen in Abschnitte, sogenannte Kompetenzniveaus, unterteilt (Hartig & Klieme, 2006; Rauch & Hartig, 2020). Für jedes Kompetenzniveau (Skalenabschnitt) wird die entsprechende Kompetenz inhaltlich beschrieben (Hartig & Klieme, 2006; Rauch & Hartig, 2020), jedoch erfolgt »innerhalb der gebildeten Kompetenzniveaus ... keine weitere inhaltliche Differenzierung der erfassten Kompetenz« (Rauch und Hartig, 2020, S. 417). Eine Veranschaulichung einer in Kompetenzniveaus unterteilten Kompetenzskala ist in Abb. 4.5 abgebildet.

Abbildung 4.5
Veranschaulichung der Unterteilung einer kontinuierlichen Kompetenzskala (Hartig, 2007, S. 87)



Im folgenden Kapitel wird erläutert, wie die Grenzen der Kompetenzniveaus bestimmt werden können.

4.2.2 Verfahren der Niveaumodellierung

Für die Entwicklung von Kompetenzniveaumodellen werden zum einen die fachbezogenen Anforderungen der Items und zum anderen die aus

der IRT-Skalierung geschätzten Itemschwierigkeiten benötigt (Rauch & Hartig, 2020). Nach Rauch und Hartig (2020) sind die Methoden zur Bestimmung der Schwellen (Grenzen) zwischen den Kompetenzniveaus ausschlaggebend für deren Entwicklung. Hierfür stehen verschiedene Methoden zur Verfügung, die unterschiedlich stark modellgeleitet sind (Hartig & Klieme, 2006). Grundsätzlich kann zwischen *a priori* und *post-hoc* Verfahren unterschieden werden (vgl. Hartig, 2007; Hartig & Klieme, 2006; Rauch & Hartig, 2020).

Sofern a-priori-Aufgabenmerkmale vorliegen, die einen Einfluss auf die Itemschwierigkeit haben, können diese zur Bestimmung der Kompetenzniveaus herangezogen werden (Hartig, 2007; Rauch & Hartig, 2020). Die aus der IRT-Skalierung resultierenden Itemschwierigkeiten der a-priori-Aufgabenmerkmale werden in Regressionsanalysen überprüft, um entsprechende Schwellenwerte auf der Kompetenzskala zu definieren (Hartig, 2007; Hartig & Klieme, 2006; Rauch & Hartig, 2020). Dieser stark modellgeleitete Ansatz wird u.a. in der Studie *Deutsch-Englisch-Schülerleistungen-International* (DESI) verwendet, das entsprechende Vorgehen wird ausführlich von Hartig (2007) beschrieben. Da im Rahmen dieser Arbeit a priori keine schwierigkeitsbestimmenden Aufgabenmerkmale vorliegen, wird auf eine ausführlichere Darstellung dieses Verfahrens verzichtet.

Bei Post-hoc-Verfahren, das gängiste wurde von Beaton und Allen (1992) vorgestellt, ist eine willkürliche Festlegung der Grenzen der Kompetenzniveaus zulässig (Hartig & Klieme, 2006). Dieses Verfahren wird anhand eines Beispiels beschrieben:

Wie bereits in Kap. 2.1.2 geschildert, werden auch die IRT skalierten Daten der TIMSS-Studie nach dem Verfahren von Beaton und Allen (1992) in ein Kompetenzniveaumodell überführt. Bei dem Post-hoc-Verfahren der Kompetenzniveaumodellierung – die Kompetenzniveaus werden also erst nach der Testdurchführung festgelegt – werden beliebige Ankerpunkte (Grenzen) auf der Testskala gewählt (Baumert et al., 1999). Nach Rauch und Hartig (2020) können die willkürlich gesetzten Ankerpunkte z.B. in gleichen Abständen oder in Relation zu den Mittelwerten bestimmter Bezugsgruppen gesetzt werden. Auf Basis dieser Ankerpunkte werden nun die Kompetenzniveaus gebildet

(Abb. 4.5), wobei diese Grenze als höchste Itemschwierigkeit des jeweiligen Kompetenzniveaus dient. Den Kompetenzniveaus werden dann diejenigen Items zugeordnet, deren Itemschwierigkeit so groß ist, dass die Probanden des entsprechenden Kompetenzniveaus die Items mit hinreichender Sicherheit lösen können (Baumert et al., 1999). Abschließend erfolgt die inhaltliche Charakterisierung eines jeden Kompetenzniveaus (Baumert et al., 1999). In der TIMSS-Studie wird dabei eine Lösungswahrscheinlichkeit von 65% als hinreichend angesehen (Baumert et al., 1999). Gleichzeitig muss aber auch die Lösungswahrscheinlichkeit der Probanden in dem nächstniedrigeren Kompetenzniveau hinreichend niedrig (<50%) sein (Beaton & Allen, 1992; Rauch & Hartig, 2020). Ein Item kann maximal einem Kompetenzniveau zugeordnet werden, es ist aber auch möglich, dass ein Item die Kriterien nicht erfüllt und daher nicht in dem Kompetenzniveaumodell berücksichtigt werden kann (Rauch & Hartig, 2020). Probanden, die ein höheres Kompetenzniveau erreichen, können auch alle Items der darunter liegenden Kompetenzniveaus mit hinreichender Sicherheit lösen. Umgekehrt können Probanden, die ein niedrigeres Kompetenzniveau erreichen, die Items eines höheren Kompetenzniveaus nicht mit hinreichender Sicherheit lösen.

Es wurde erwähnt, dass die Ankerpunkte die Grenzen des Kompetenzniveaus und damit die höchste Itemschwierigkeit des entsprechenden Kompetenzniveaus darstellen. Hartig (2007) variiert dieses Vorgehen und spricht von der »Nachbarschaft der Schwellen« (Hartig, 2007, S. 87), die in Abb. 4.5 mit den Ellipsen um die Items dargestellt ist. Die inhaltliche Beschreibung eines Kompetenzniveaus, d.h. der Probanden deren Personenfähigkeit innerhalb des Kompetenzniveaus liegt, erfolgt anhand der Items, die am Anfang des jeweiligen Kompetenzniveaus positioniert sind (Hartig, 2007). Explizit weist Hartig (2007) darauf hin, dass die Items innerhalb eines Kompetenzniveaus für die inhaltliche Beschreibung ungeeignet sind, da die Items an der oberen Schwelle »eher charakteristisch für die Leistungen von Schülern [Probanden] auf dem nächsthöheren Niveau« (Hartig, 2007, S. 87) sind. In diesem Zusammenhang spricht Hartig (2007) nicht von einer strikten Grenze zwischen den Kompetenzniveaus, sondern verwendet den Begriff der Schwelle.

Auch in PISA kommt das Post-hoc-Verfahren zur Anwendung, in dieser Studie werden die ermittelten Itemschwierigkeiten zur Bestimmung der Schwellenwerte der Kompetenzniveaus herangezogen, teilweise erfolgt bereits im Vorfeld der Datenerhebung eine Einordnung der Items in die erwarteten Kompetenzniveaus (Hartig & Klieme, 2006).

Mit den in diesem Kapitel geschaffenen methodischen Grundlagen erfolgt die IRT-Skalierung in Kap. 6.2 und die Entwicklung der Kompetenzniveaumodelle in Kap. 6.3.

Kapitel 5

Untersuchungsdesign

Um die in Kap. 3 beschriebenen Forschungsfragen beantworten und die Hypothesen bewerten zu können, wurden im vorangegangenen Kapitel (Kap. 4) die methodischen Grundlagen gelegt. Des Weiteren ist ein geeignetes Untersuchungsdesign notwendig, welches in diesem Kapitel beschrieben wird. Dabei wird ausschließlich auf die Studie FUNDAMENT Bezug genommen, sofern nicht anders angegeben. Das Untersuchungsdesign von ALSTER kann in Dammann und Lang (2018, 2019) und J. Müller et al. (2018) nachgelesen werden.

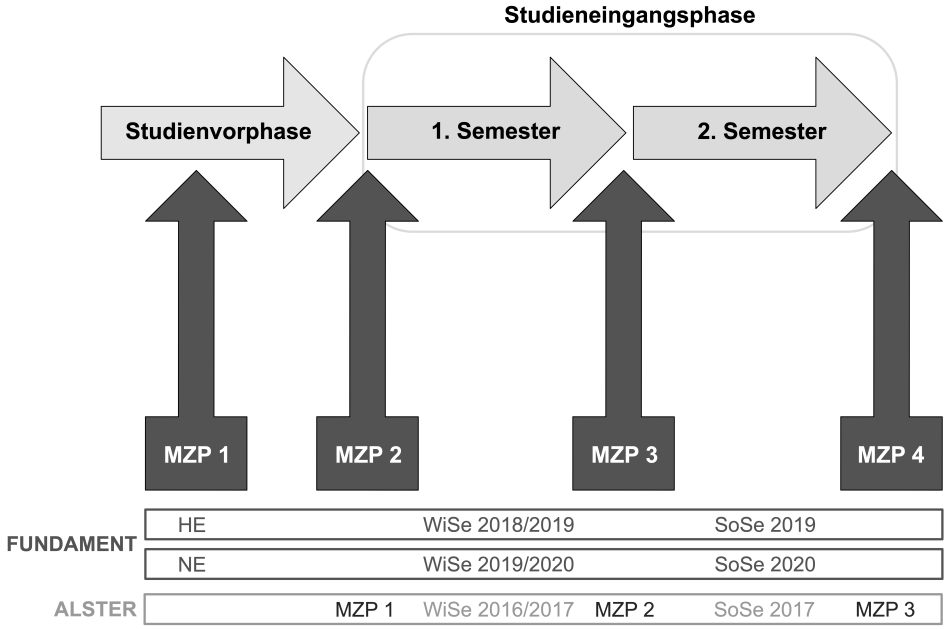
Die Stichprobe wird anhand der erhobenen demographischen Daten deskriptiv beschrieben, wobei Kennwerte des Gesamtdatensatzes dargestellt und Unterschiede zwischen den beiden Studien und den beiden Standorten herausgearbeitet werden.

5.1 Beschreibung der Untersuchung

Wie bereits in Kap. 2.5.1.1 und 2.5.1.2 kurz beschrieben, basiert das Längsschnittdesign von FUNDAMENT auf vier MZP. Während sich der erste MZP in der Studienvorphase befindet, sind die drei weiteren MZP in der Studieneingangsphase in den ersten beiden Fachsemestern angesiedelt (Abb. 5.1). Um eine größere Probandenanzahl zu generieren, wurden die Daten in zwei Kohorten erhoben. In der Haupterhebung (HE) wurden Probanden untersucht, die im WiSe 2018/2019 ihr erstes Fachsemester im Bauingenieurwesen an der UDE begonnen haben (Abb. 5.1).

Dementsprechend liegt der erste MZP vor Beginn des WiSe 2018/2019, MZP 2 zu Beginn und MZP 3 am Ende des Semesters. Der vierte MZP findet am Ende des SoSe 2019 statt, in dem die Probanden ihr zweites Fachsemester abschließen. Bei der Nacherhebung (NE), der zweiten untersuchten Kohorte, handelt es sich ebenfalls um Studienanfänger des Bauingenieurwesens an der UDE, die jedoch ein Jahr später, im WiSe 2019/2020, ihr erstes Fachsemester begonnen haben (Abb. 5.1). Die MZP sind analog zu denen der Haupterhebung gesetzt, jedoch um ein Jahr nach hinten verschoben.

Abbildung 5.1
FUNDAMENT/ALSTER - Erhebungsdesign



Anmerkungen. MZP = Messzeitpunkt; HE = Haupterhebung; NE = Nacherhebung; WiSe = Wintersemester; SoSe = Sommersemester.

Das Testinstrument des ersten MZP hat keine inhaltliche Verankerung mit den anderen drei MZP und ist durch die Verwendung von OSA 1 und OSA 2 in sich abgeschlossen. Die Erhebung an MZP 1 wurde ausschließlich online durchgeführt. Die an MZP 1 in den beiden OSA erzielten Ergebnisse werden daher separat betrachtet (Kap. 6.4.3). Die weiteren MZP sind dagegen miteinander verankert, wobei der MZP 2 den Ausgangspunkt zu Beginn des Studiums darstellt, während die Test-

instrumente der beiden anderen MZP am Ende des ersten (MZP 3) bzw. des zweiten (MZP 4) Fachsemesters eingesetzt werden. Mit Semesterende ist in diesem Zusammenhang eine zeitliche Nähe zu den Klausurterminen gemeint, da nach Dammann (2016) davon ausgegangen werden kann, dass die entsprechenden Kompetenzausprägungen zu diesem Zeitpunkt sehr hoch sind. An den drei MZP wurden die fachlichen Leistungstests Technische Mechanik und Rechenfähigkeit als Paper & Pencil-Tests eingesetzt und auch die erzielten Klausurnoten erhoben. Abweichend davon wurden die fachlichen Leistungstests bei der Nacherhebung an MZP 4 nicht als Paper & Pencil-Tests appliziert, sondern in Form einer LimeSurvey Onlinebefragung (LimeSurvey GmbH, o. J.). Dies hat den Hintergrund, dass im SoSe 2020 aufgrund der COVID-19-Pandemie (Dittler & Kreidl, 2021) keine Veranstaltungen vor Ort stattfanden und somit auch die Testung nicht entsprechend durchgeführt werden konnte. Neben den fachlichen Leistungstests wurden auch demographische Variablen erfasst. Diese wurden sowohl an MZP 1 als auch an MZP 2 erhoben. Für den Fall, dass Probanden nur MZP 3 und/oder MZP 4 absolvierten, wurde die Möglichkeit einer nachträglichen Erhebung der demographischen Variablen angeboten. Zusätzliche Testinstrumente zur Erhebung der kognitiven Grundfähigkeit, des typischen intellektuellen Engagements oder des beruflichen Interesses wurden ebenfalls eingesetzt, jedoch von zu wenigen Probanden vollständig bearbeitet, daher wird auf eine Auswertung an dieser Stelle verzichtet.

In der Haupterhebung wurden die Klausuren der Veranstaltungen des ersten (Technische Mechanik 1 und Mathematik 1) und zweiten Fachsemesters (Technische Mechanik 2 und Mathematik 2) jeweils als zwei Teilklausuren in der Mitte und am Ende des jeweiligen Semesters gestellt. In der Nacherhebung gab es eine Änderung der Prüfungsordnung bezüglich der beiden Lehrveranstaltungen der Technischen Mechanik. Seit WiSe 19/20 werden keine Teilklausuren mehr angeboten, sondern nur noch jeweils eine Abschlussklausur. Die entsprechenden Klausurnoten wurden für die beiden Veranstaltungen der Technischen Mechanik bis WiSe 20/21 semesterweise erhoben. Somit sind auch die Noten der nachgeschriebenen Teilklausuren bzw. Klausuren im Datensatz enthalten. Falls Probanden mehrere Versuche benötigten, um eine Prüfung

zu bestehen, wird in den Auswertungen (siehe Kap. 6) nur die beste erreichte Note berücksichtigt. Aus den beiden Noten der Teilprüfungen eines Fachsemesters wird ein Mittelwert gebildet, um eine vergleichbare Endnote zu erhalten. Daher kann es vorkommen, dass diese Noten von den üblichen Drittelnoten abweichen. Gleiches gilt für die Noten der beiden Mathematikveranstaltungen. Aus organisatorischen Gründen wurden die Prüfungsergebnisse jedoch nur einmal am Ende der Haupterhebung erhoben. Dies bedeutet, dass für die Probanden der Nacherhebung (Studienbeginn WiSe 19/20) nur die Klausurergebnisse der Technischen Mechanik und nicht die der Mathematik vorliegen.

ALSTER verwendet eine andere Nummerierung der MZP, da in dieser Studie keine Erhebung in der Studienvorphase durchgeführt wurde und auch keine iOM zum Einsatz kamen (Abb. 5.1). Die Nummerierung wurde jedoch zur besseren Übersicht an die von FUNDAMENT angepasst. In ALSTER wurden Studierende im ersten Fachsemester des Bauingenieurwesens sowohl an der UDE als auch an der RUB des WiSe 2016/2017 untersucht. An allen MZP wurden die Daten mit Hilfe von Paper & Pencil-Tests erhoben. Die in ALSTER an MZP 4 erhobenen Daten zur Rechenfähigkeit werden aufgrund von Kodierungsproblemen in dieser Arbeit nicht berücksichtigt.

5.2 Testitems

Zunächst findet eine Beschreibung des Fragebogens zur Erhebung der demographischen Daten statt, danach die Beschreibung der Items, die an MZP 1 im OSA eingesetzt wurden. Abschließend erfolgt die Beschreibung der Items der fachlichen Leistungstests der Technischen Mechanik (Fachwissen und Modellierungsfähigkeit) und der Rechenfähigkeit.

Der Fragebogen zu den demographischen Variablen orientiert sich an dem von ALSTER verwendeten Fragebogen, weshalb eine Vergleichbarkeit zwischen den beiden Studien gewährleistet ist. Neben allgemeinen Variablen wie u.a. Geschlecht, Geburtsjahr, Muttersprache werden auch Variablen zur Bestimmung der Bildungsherkunft erhoben. Die schulische Bildung wird mit Fragen zur Hochschulzugangsberechtigung (u.a. Abschlussnote, Kurswahl in der Sekundarstufe II, (Bundes-)Land in dem

die Hochschulreife erworben wurde, Abschlussnote (in Punkten) verschiedener Fächer in der Oberstufe) erfasst, während auch eine vorherige Berufsausbildung oder Studienerfahrung abgefragt wird. Die Probanden werden auch gefragt, ob sie im Vorfeld des Studiums an einem Vorkurs teilgenommen haben, und wenn ja, um welche Art von Vorkurs es sich handelte.

Im OSA - an MZP 1 - sind die Inhaltsbereiche der naturwissenschaftlichen (NG) und mathematischen Grundlagen (MG) abgebildet (Kap. 2.3.1). Das OSA wurde ausschließlich in der Studie FUNDAMENT verwendet und stellt hier den ersten MZP dar. Die in FUNDAMENT entwickelten Items wurden alle zweimal eingesetzt und können daher alle als Anker innerhalb des MZP betrachtet werden: vor Absolvierung des OV (OSA 1) und nochmals nach Bearbeitung von Teilen oder des gesamten OV (OSA 2). Die Items der naturwissenschaftlichen Grundlagen umfassen die Themen Kräfte/Momente, Schwerpunkt, Lager, Stabilität und Zug, Druck, Biegung. Die mathematischen Grundlagen decken die Themen Bruchrechnung, quadratische Funktionen, Gleichungssysteme, Trigonometrie, Vektorrechnung und Analysis ab. Die genaue Verteilung der Items kann in Tab. 5.1 eingesehen werden. Für weitere Details siehe Kap. 2.5.1.1.

Tabelle 5.1
Verteilung der eingesetzten Items an MZP 1

	NG	MG	n_{Items}
OSA 1	12	23	35
OSA 2	12	23	35

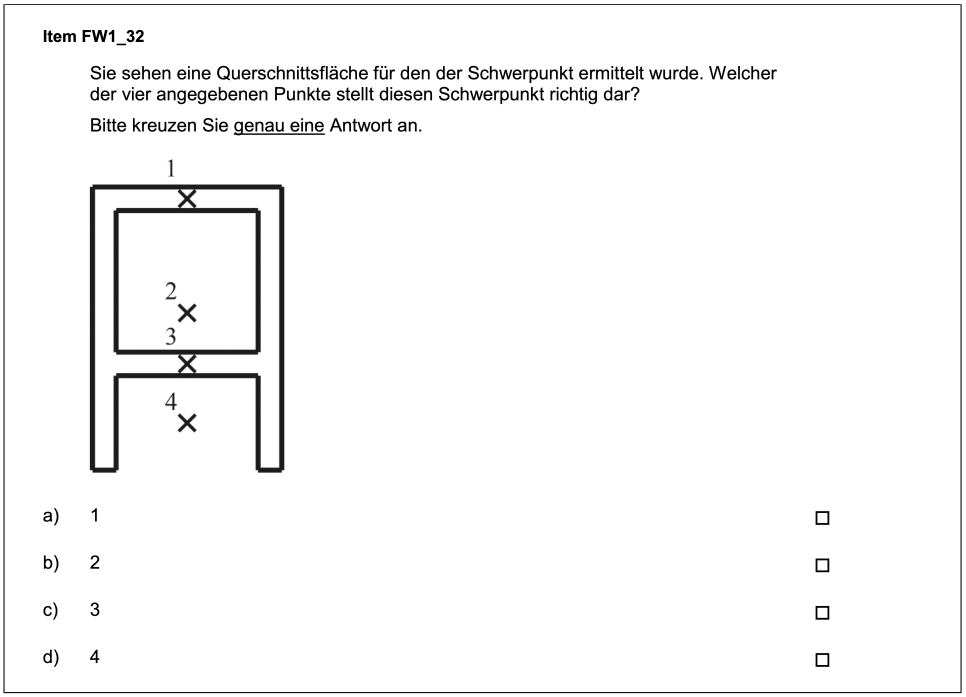
Anmerkungen. NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; n_{Items} = Summe Anzahl der Items; OSA = Online-Self-Assessment.

Die anderen drei MZP (MZP 2-4) basieren auf den in ALSTER entwickelten fachlichen Leistungstests (Dammann & Lang, 2018), ergänzt durch das in FUNDAMENT entwickelte Testinstrument für die Technische Mechanik 2 (TM_4). Während die Testinstrumente zum Fachwissen (FW), der Modellierungsfähigkeit (MF) und die Technische Mechanik 2-Items an MZP 4 (TM_4) den Inhaltsbereichen der Technischen Mechanik zugeordnet sind, befasst sich das letztgenannte Testin-

strument mit der Rechenfähigkeit (RF), also mathematischen Themen. Die von ALSTER adaptierten Testinstrumente wurden nicht vollständig übernommen, sondern es wurde aus testökonomischen Gründen eine Vorauswahl getroffen.

Wie bereits in Kap. 2.5.2 beschrieben, wurden bei der Konstruktion der Items zum Fachwissen die Facetten der Komplexitätsdimension berücksichtigt. Als Beispiel ist in Abb. 5.2 ein Item abgebildet, das am zweiten MZP das Fachwissen zum Themenbereich ›Schwerpunkt‹ abfragt. Zur Beantwortung der Fragestellung wird das Fachwissen über die Lage eines Flächenschwerpunktes benötigt.

Abbildung 5.2
Item von MZP 2 zum Fachwissen des Themengebiets ›Schwerpunkt‹



Die Modellierungsfähigkeit der Probanden wird durch Items zu den drei relevanten Modellierungsschritten (Mathematisieren/Konzeptualisieren, Abstrahieren und fachlich Interpretieren) erfasst (siehe Kap. 2.5.2). Abb. 5.3 zeigt ein Item des dritten MZP zum Themengebiet ›Tragwerke‹ des zweiten Modellierungsschritts (Mathematisieren/ Konzeptualisieren). Der Operator in diesem Beispiel ist das ›Freischnei-

den der Kräfte eines mechanischen Systems‹. Erst wenn das Freischneiden der Kräfte der dargestellten Aufgabe erfolgt ist, können die für das Momentengleichgewicht benötigten Variablen bestimmt werden.

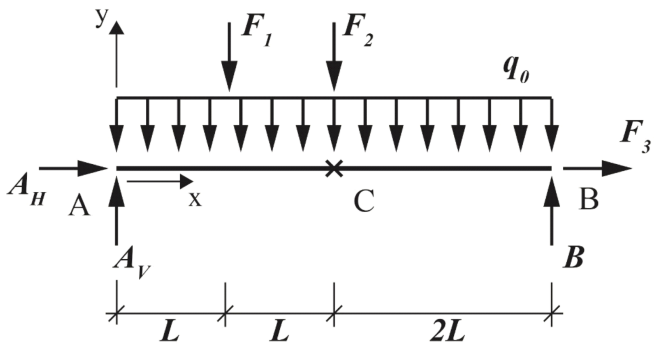
Abbildung 5.3
Item von MZP 3 zur Modellierungsfähigkeit (Modellierungsschritt 2) des Themengebiets ›Tragwerke‹

MF2_2_27

Sie sehen ein ebenes Tragwerk, bei dem für den Punkt C das Momenten-Gleichgewicht nachgewiesen werden soll. Welche der gegebenen Größen müssen für diesen Nachweis in diesem Fall unbedingt berücksichtigt werden?

Bitte kreuzen Sie genau eine Antwort an.

Geg.: $F_1, F_2, F_3, q_0, A_H, A_V, B, L$



a) A_V, B, F_1, q_0, L ☐

b) A_V, B, F_1, F_2, q_0, L ☐

c) $A_V, B, F_1, F_2, F_3, q_0, L$ ☐

d) A_V, B, F_1, L ☐

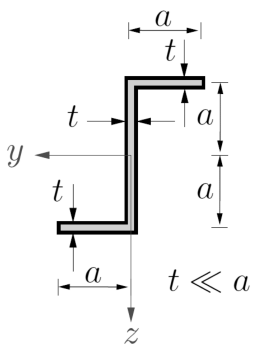
Da in den adaptierten ALSTER-Items zwei Themengebiete der Technischen Mechanik 2 nicht berücksichtigt sind, wurden in FUNDAMENT Items zu den Themenbereichen Spannungszustand und Balken entwickelt. Die TM_4-Items bilden ein eigenständiges Testinstrument und enthalten sowohl Items zum Fachwissen als auch zur Modellierungsfähigkeit. Dieses Testinstrument wurde ausschließlich in FUNDAMENT eingesetzt, ALSTER-Daten liegen hierzu nicht vor. Abb. 5.4 zeigt eines dieser TM_4-Items, welches der Modellierungsfähigkeit zugeordnet wer-

den kann. Hier wird der Modellierungsschritt 3 getestet, bei dem das Abstrahieren im Vordergrund steht und der Operator ›Aufstellen der fachlich-mathematischen Gleichungen‹ zum Tragen kommt.

Abbildung 5.4
Item von MZP 4 zur Technischen Mechanik 2 zum Themengebiet ›Balken‹

Item TM2_07

Welche Aussage trifft für die axialen Flächenträgheitsmomente zu?



Bitte kreuzen Sie genau eine Antwort an.

a) $I_y = I_z$ ☐

b) $I_y > I_z$ ☐

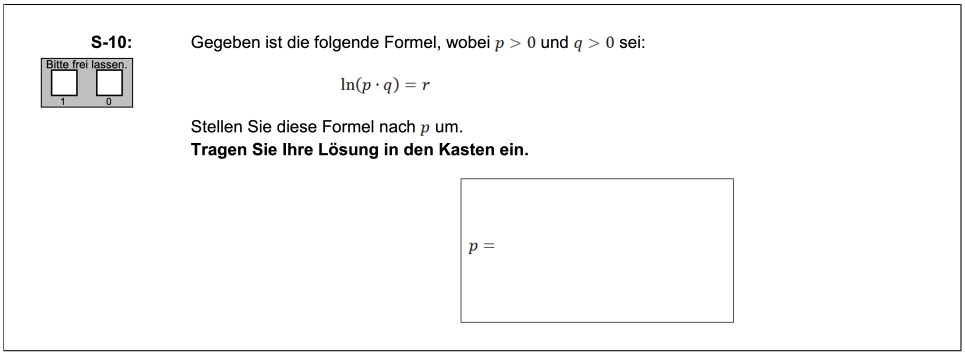
c) $I_z = 0$ ☐

d) $I_y < 0$ ☐

Neben den genannten Items, die sich auf mechanische Inhalte beziehen, beschränken sich die Items zur Rechenfähigkeit auf grundlegende mathematische Inhaltsbereiche. Mit diesem Testinstrument wird das mathematische Schulwissen zu den Inhaltsbereichen Rechnen, Terme und Gleichungen, Trigonometrie, Differenzieren und Integrieren erfasst. Es werden also nicht die Inhalte der Mathematikveranstaltungen des Studiums geprüft. Das in ALSTER fächerübergreifend eingesetzte Testinstrument zur Erfassung der Rechenfähigkeit besitzt ein offenes Antwortformat (Abb. 5.5).

Dieses wurde von Dammann für einen Rechenfähigkeitstest für Studi-

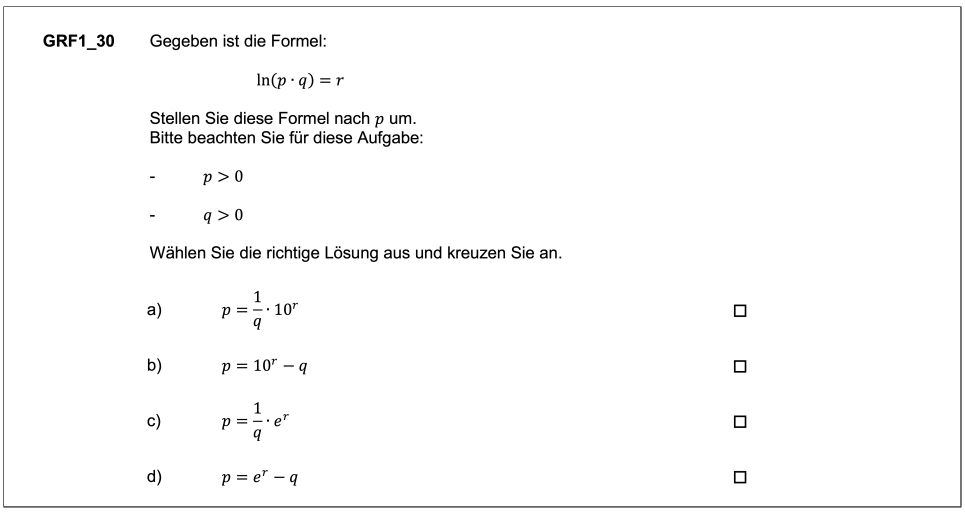
Abbildung 5.5
Item von MZP 2 zur Rechenfähigkeit (ALSTER)



enanfänger an der Hochschule Furtwangen in ein geschlossenes Antwortformat – Multiple-Choice mit vier Antwortmöglichkeiten – umgewandelt (Abb. 5.6). Das überarbeitete Testinstrument von Dammann wurde in FUNDAMENT zur Erfassung der Rechenfähigkeit eingesetzt.

Ob die unterschiedlichen Antwortformate einen Einfluss auf die erzielten Ergebnisse haben, wird bei der Skalierbarkeit des Gesamtdatensatzes überprüft (Kap. 6.2.1). Aufgrund von Kodierungsproblemen werden RF-Daten von ALSTER an MZP 4 nicht berücksichtigt.

Abbildung 5.6
Item von MZP 2 zur Rechenfähigkeit (FUNDAMENT)



5.3 Testheftdesign

Da im vorangegangenen Kapitel (5.2) bereits der Aufbau und die Testung an MZP 1 beschrieben wurde, wird in diesem Kapitel lediglich das Testheftdesign der MZP 2-4 erläutert.

In der Studie FUNDAMENT wurden sowohl in der Haupterhebung, als auch in der Nacherhebung die gleichen Items eingesetzt, es wird ein klassisches A/B Testheftdesign verwendet. Während im Testheft A die Testinstrumente in der Reihenfolge Fachwissen, Modellierungsfähigkeit, (TM_4) und Rechenfähigkeit – innerhalb der Testinstrumente orientiert sich die Reihenfolge der Items an der ursprünglichen Item-Bezeichnung – auf jeweils einer Seite angeordnet sind, wird im Testheft B die Reihenfolge verändert: Bei der Haupterhebung an MZP 2 sind die Items innerhalb der Testinstrumente ungeordnet, bei MZP 3 ist die Anordnung entgegengesetzt zu Testheft A, aber die Reihenfolge der Testinstrumente bleibt erhalten. Bei MZP 4 sind sowohl die Reihenfolge der Testinstrumente als auch die Anordnung der Items innerhalb der Testinstrumente vertauscht. Die Items von MZP 2 und 3 der Nacherhebung werden innerhalb der Testinstrumente in umgekehrter Reihenfolge angeordnet. MZP 4 der Nacherhebung hat aufgrund der Online-Befragung kein A/B Testheftdesign.

An MZP 3 waren die Lösungsalternativen des Items *MF3_5_08* in der Haupterhebung fehlerhaft, daher wurde das Item an diesem MZP für beide Erhebungen nachträglich aus dem Datensatz entfernt. Ohne dieses Item ergibt sich die in Tab. 5.2 dargestellte Verteilung der Items.

Um Probanden verschiedener Testformen (z.B. MZP oder Studien) hinsichtlich ihrer IRT geschätzten Personenfähigkeit vergleichen zu können, sind Ankeritems notwendig (Rauch & Hartig, 2020). Diese in den verschiedenen Testformen verwendeten Ankeritems werden »auf einer Skala mit einer gemeinsamen Metrik« (Rauch und Hartig, 2020, S. 412) verankert.

Um die Skalierung auf eine gemeinsame Metrik zu gewährleisten, müssen sowohl die MZP innerhalb von FUNDAMENT und ALSTER als auch die Studien FUNDAMENT und ALSTER untereinander verankert werden. Die Anzahl der Ankeritems ist in Tab. 5.3 aufgelistet. Da die TM_4-Items nur an einem MZP und in einer Studie appliziert

Tabelle 5.2
Verteilung der eingesetzten Items an MZP 2-4

	FW	MF	TM_4	RF	n_{Items}
FUNDAMENT					
MZP 2	25	18		25	68
MZP 3	16	29		14	59
MZP 4	16	19	17	14	70
ALSTER					
MZP 2	40	47		27	114
MZP 3	50	68		30	148
MZP 4	50	68			148

Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4; RF = Rechenfähigkeit; n_{Items} = Summe Anzahl der Items; MZP = Messzeitpunkt.

Tabelle 5.3
Verteilung der Ankeritems an MZP 2-4

	FW	MF	RF
FUNDAMENT			
MZP 2 → MZP 3	7	7	14[a]
MZP 3 → MZP 4	14	19[a]	14[a]
ALSTER			
MZP 2 → MZP 3	33	41	27[a]
MZP 3 → MZP 4	50[a]	68[a]	
ALSTER ↔ FUNDAMENT			
MZP 2	25[a]	18[a]	11
MZP 3	16[a]	29[a]	7
MZP 4	16[a]	19[a]	

Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4; RF = Rechenfähigkeit; MZP = Messzeitpunkt.
[a] Alle eingesetzten Items sind Ankeritems.

wurden, sind sie keine Ankeritems und werden daher in der Tabelle nicht aufgeführt.

Die Tabelle 5.3 ist so zu lesen, dass z.B. in FUNDAMENT sieben der an MZP 2 eingesetzten FW-Items auch an MZP 3 verwendet wurden und somit Ankeritems sind. Bei den RF-Items wurden beispielsweise 14 Items von MZP 2 auch an MZP 3 eingesetzt, da aber insgesamt nur 14 RF-Items an diesem MZP appliziert wurden (siehe Tab. 5.2), sind alle RF-Items Ankeritems. Eine ähnliche Situation zeigt sich bei der Verankerung zwischen ALSTER und FUNDAMENT. Hier sind alle in FUNDAMENT eingesetzten FW- und MF-Items Ankeritems.

5.4 Testdurchführung

Bei der Beschreibung der Testdurchführung ist wie bisher zwischen MZP 1 und den MZP 2-4 zu unterschieden. Für alle MZP gilt jedoch, dass die Probanden einer Einwilligungserklärung zustimmen mussten, damit die Testergebnisse, die Nutzungszahlen der Online-Elemente und die Klausurnoten gespeichert und ausgewertet werden durften. Es wurde klar kommuniziert, dass die Teilnahme an der Studie FUNDAMENT (an einzelnen oder allen MZP) freiwillig ist. Bei der Bearbeitung der Testinstrumente sind keine Hilfsmittel wie Taschenrechner, Vorlesungsskripte oder Formelsammlungen zugelassen. Die erreichten Probandenzahlen sind in Tab. 5.4 aufgeführt.

Tabelle 5.4
Probandenzahlen FUNDAMENT

	HE	NE	$n_{Probanden}$
MZP 1			
OSA 1	20	11	31
OSA 2	12	2	14
MZP 2	180	116	296
MZP 3	63	53	116
MZP 4	91	14	105

Anmerkungen. HE = Haupterhebung; NE = Nacherhebung; $n_{Probanden}$ = Summe Anzahl der Probanden; MZP = Messzeitpunkt; OSA = Online-Self-Assessment.

Da sich der erste MZP in der Studienvorphase befindet, können die Probanden die Testinstrumente nicht vor Ort bearbeiten. Daher werden die Testinstrumente in der Moodle-Umgebung der UDE in Kombination mit JACK bereitgestellt. Über ›Vorausssetzungsregeln‹ in Moodle wird festgelegt, dass die Testinstrumente in einer bestimmten Reihenfolge bearbeitet werden: Fragebogen zu demographischen Variablen, OSA naturwissenschaftliche Grundlagen und OSA mathematische Grundlagen. Erst wenn eines der Testinstrumente abgeschlossen ist, wird das nächste freigeschaltet. Nach Abschluss der beiden OSA erhalten die Probanden eine Rückmeldung zu beiden Themenbereichen mit Handlungsempfehlungen zur Auffrischung und Erweiterung ihres Wissens im OV. Nach Abschluss aller drei Testinstrumente wird das OSA 2 naturwissenschaftliche Grundlagen freigeschaltet, nach dessen Abschluss das OSA 2 mathematische Grundlagen. Diese beiden OSA dienen der Überprüfung der Wirksamkeit des OV und sollten dementsprechend erst bearbeitet werden, wenn Teile oder der gesamte OV bearbeitet wurde (siehe Kap. 2.5.1.1). Ob Hilfsmittel verwendet wurden oder ob z.B. gemeinsam mit anderen an einem OSA gearbeitet wurde, kann aufgrund der begrenzten administrativen Möglichkeiten der Online-Befragung nicht geklärt werden, die Ergebnisse deuten jedoch nicht darauf hin.

Die Probandenakquise in der Studienvorphase erwies sich als sehr problematisch. Studieninteressierte, die zum Zeitpunkt des Mailversands im Studiengang Bauingenieurwesen an der UDE immatrikuliert waren, konnten nur per E-Mail vom Einschreibewesen über die Verfügbarkeit von OSA und OV informiert und um Teilnahme gebeten werden. In dieser E-Mail wurde auch auf die ausgelobte Probandenvergütung (10 EUR für die Teilnahme an OSA 1 und 20 EUR für den erfolgreichen Abschluss von OSA 2) hingewiesen. Allerdings konnten von insgesamt 441 per Mail angeschriebenen Studieninteressierten nur knapp 14% zur Teilnahme bewegt werden. Hierbei handelt es sich allerdings nicht um vollständige Teilnahmen, sondern überwiegend wurde nur der Fragebogen zu den demographischen Variablen ausgefüllt. Insgesamt beendeten 31 Probanden den OSA 1, während nur 14 den OSA 2 vollständig absolviert haben (siehe Tab. 5.4). Es zeigt sich, dass die Aufwandsentschädigung für die Probanden (insgesamt 30 EUR) keinen ausreichenden Anreiz zur

Teilnahme darstellt. Auch die Kommunikation mit den Studierenden vor Studienbeginn ist ungünstig, da nur die Universitäts-E-Mail-Adressen zur Verfügung stehen und es fraglich ist, ob die Studieninteressierten diese bereits Wochen vor Studienbeginn regelmäßig nutzen.

Die Durchführung der anderen drei MZP war ursprünglich ebenfalls online (zeit- und ortsunabhängig) geplant, wurde aber aufgrund der sehr geringen Probandenzahlen während der Pilotierung für die Haupt- und Nacherhebung auf Paper & Pencil Tests in Präsenz umgestellt. Für die Bearbeitung der applizierten Testinstrumente (siehe Kap. 5.3) standen jeweils 90 Minuten zur Verfügung. Um eine möglichst große Probandenzahl zu erreichen, wurde der zweite MZP in die erste Veranstaltung der Technischen Mechanik 1 gelegt. Die MZP 3 und 4 fanden dagegen in der letzten Vorlesungswoche im Anschluss an eine Hörsaalübung in der eigentlichen Vorlesungszeit der Technischen Mechanik 1 bzw. 2 statt. Das selbstständige Arbeiten der Probanden konnte durch den Sitzabstand zwischen den Probanden, das A/B Testheftdesign und dem Aufsichtspersonal gewährleistet werden. Wie bereits in Kap. 5.1 erwähnt, wurde der vierte MZP aufgrund der COVID-19-Pandemie in der Nacherhebung online durchgeführt. Die Bearbeitungszeit wurde auch in LimeSurvey-Befragung auf 90 Minuten begrenzt. Eine Überprüfung des Einsatzes von Hilfsmitteln war nicht möglich, die Ergebnisse zeigen jedoch keine entsprechenden Auffälligkeiten.

Auch an den MZP 2-4 wurde eine Probandenvergütung ausgelobt (MZP 2 und 3 je 20 EUR, MZP 4 30 EUR). Um die Probanden zu einer gewissenhaften Bearbeitung des Tests zu motivieren, wurde zusätzlich unter den besten 10% eines jeden MZP ein Semesterbeitrag verlost. Wie schon bei MZP 1 stellt die Probandenvergütung auch bei den späteren MZP keinen ausreichenden Anreiz dar. Insbesondere im Längsschnitt ist ein starker Schwund zu beobachten (siehe Tab. 5.4). In der Nacherhebung wurden zusätzlich Bonuspunkte für die Klausuren der Technischen Mechanik angeboten, um den Anreiz für die Studierenden zu erhöhen. Dennoch konnten absolut nur weniger Probanden erreicht werden als in der Haupterhebung, dies ist u.a. auf die um 22% geringere Grundgesamtheit der Studierenden im ersten Fachsemester zurückzuführen (siehe Tab. 5.4). Die Online-Befragung in der Nacherhebung an

MZP 4 zeigte ähnliche Probleme der Probandenakquise wie an MZP 1. Zwar konnten die Probanden über den Moodle-Kurs der Technischen Mechanik 2 mehrfach gezielt an die Teilnahme erinnert werden, dennoch blieben die Zahlen deutlich hinter der Paper & Pencil Testung in der Haupterhebung zurück (siehe Tab. 5.4).

5.5 Stichprobenbeschreibung

In diesem Kapitel soll die Stichprobe anhand der erhobenen demographischen Daten deskriptiv beschrieben werden. Die erreichte Stichprobe setzt sich aus Studierenden des Studiengangs Bauingenieurwesen der UDE und der RUB zusammen. Wie bereits in Kap. 5.1 beschrieben, stammen die Daten aus den Erhebungen der Forschungsprojekte FUNDAMENT und ALSTER. Während die FUNDAMENT-Daten ausschließlich an der UDE erhoben wurden – zwei Kohorten von Studierenden im ersten Fachsemester im WiSe 2018/2019 (Haupterhebung) bzw. 2019/2020 (Nacherhebung) – stammen die ALSTER-Daten sowohl von der UDE als auch von der RUB aus demselben Semester (WiSe 2016/ 2017).

Insgesamt liegen für 546 Datensätze demographische Daten vor, die jedoch hinsichtlich der Bearbeitung der Testinstrumente an den einzelnen MZP nicht vollständig sind. So haben einige Probanden nur an einzelnen MZP teilgenommen. Darüber hinaus liegen weitere 56 Datensätze vor, die zwar Ergebnisse zu den einzelnen Testinstrumenten an verschiedenen MZP enthalten, diese Probanden haben jedoch den Fragebogen zu den demographischen Variablen nicht bearbeitet und werden daher bei der Beschreibung der Stichprobe nicht berücksichtigt.

Im Folgenden werden die demographischen Daten der Gesamtstichprobe sowie signifikante Unterschiede zwischen den beiden Studien (FUNDAMENT und ALSTER) und den beiden Standorten (UDE und RUB) dargestellt. Zur Ermittlung signifikanter (zweiseitiger) Unterschiede werden der Chi-Quadrat-Test bzw. der exakte Fisher-Test (kategoriale Daten) und der t-Test (intervallskalierte Daten) für unabhängige Stichproben verwendet. Die Größe dieser Unterschiede bezogen auf die Gesamtpopulation wird durch die Effektstärke beschrieben (Cohen,

1988; Eid et al., 2017). Die Effektstärke wird je nach Test mit Cohen's d, dem Phi-Koeffizitent (φ) oder Cramer's V angegeben und jeweils in drei Abstufungen interpretiert (Cohen, 1988). Als Effektstärkenmaß des Chi-Quadrat-Tests bzw. des exakten Fisher-Tests für unabhängige Stichproben werden φ bzw. V herangezogen, während φ bei dichotomen kategorialen Variablen (2x2 Kreuztabellen) angewendet wird, kommt V zur Anwendung, wenn mindestens eine Variable mehr als zwei Ausprägungen hat (Sedlmeier & Burkhardt, 2021). Die Abstufungen der beiden Effektstärken lautet:

- $|\varphi|$ bzw. $|V| \geq 0.1$ - kleiner Effekt
- $|\varphi|$ bzw. $|V| \geq 0.3$ - mittlerer Effekt
- $|\varphi|$ bzw. $|V| \geq 0.5$ - großer Effekt (Cohen, 1988).

Die Interpretation des t-Tests für unabhängige Stichproben erfolgt mit Hilfe des Effektstärkemaßes d in folgender Abstufung:

- $|d| \geq 0.2$ - kleiner Effekt
- $|d| \geq 0.5$ - mittlerer Effekt
- $|d| \geq 0.8$ - großer Effekt (Cohen, 1988).

Die Tests werden mit einem Signifikanzniveau von $\alpha = .05$ durchgeführt, die Berechnung erfolgt mit der Statistiksoftware ›R‹ (R Core Team, 2023) und den ›R‹-Paketen ›effectsize‹ (Ben-Shachar et al., 2020) und ›psych‹ (Revelle, 2023).

Die Geschlechterverteilung und die Ausschöpfungsquoten sind in Tab. 5.5 dargestellt. Im Gesamtdatensatz von 546 Probanden sind 188 Frauen enthalten, was einem Frauenanteil von 34% entspricht. Zum Vergleich: In den Bachelor-Studiengängen des Bauingenieurwesens an der UDE sind derzeit 849 Studierende eingeschrieben, davon knapp 37% Frauen. Der Frauenanteil ist auffallend hoch, üblicherweise liegt er in ingenieurwissenschaftlichen Studiengängen zwischen 12% und 19% (Klöpping et al., 2017).

Tabelle 5.5
Stichprobe - Geschlechterverteilung und Ausschöpfungsquote

	männlich	weiblich	n	N	%
		FUNDAMENT			
UDE - WiSe 2018/2019	131	67	198	319	62.1
UDE - WiSe 2019/2020	93	49	142	249	57.0
Gesamt (FUNDAMENT)	224	116	340		
		ALSTER			
UDE - WiSe 2016/2017	60	30	90	332	27.1
RUB - WiSe 2016/2017[a]	74	42	116		
Gesamt (ALSTER)	134	72	206		
$\sum$	358	188	546		

Anmerkungen. n = Stichprobengröße; N = Grundgesamtheit der Studierenden im ersten Fachsemester; % = Ausschöpfungsquote; UDE = *Universität Duisburg-Essen*; WiSe = Wintersemester; RUB = *Ruhr-Universität Bochum*; $\sum$ = Summe aller Probanden.
[a] Zahl der Studierenden im ersten Fachsemester nicht verfügbar.

Die Ausschöpfungsquote der beiden Kohorten von FUNDAMENT liegt mit 57% bzw. 62%, deutlich über der von ALSTER (27%). Die Ausschöpfungsquote für die ALSTER-Erhebung an der RUB konnte nicht berechnet werden, da die Zahl der Studierenden im ersten Fachsemester nicht vorliegt.

Das Durchschnittsalter der Gesamtstichprobe beträgt 20.5 Jahre ($SD = 2.8$). Es zeigt sich ein signifikanter Unterschied mit einem kleinen Effekt zwischen den beiden Studien, wobei die Probanden der Studie FUNDAMENT im Durchschnitt 0.7 Jahre ($SD = 1.2$) älter sind ($95\% - CI[0.27, 1.13]$), $t(543) = 3.18$, $p = .002$, $d = 0.27$).

52% der Probanden geben Deutsch als ihre Muttersprache an. Auch hier zeigt sich ein signifikanter Unterschied zwischen den beiden Studien ($\chi^2(1) = 4.32$, $p = .038$, $\varphi = 0.08$). Während bei ALSTER 58% der Probanden Deutsch als ihre Muttersprache angeben, sind es bei FUNDAMENT weniger als die Hälfte (48%). 68% der Nicht-Muttersprachler schätzen ihre Deutschkenntnisse auf einer dreistufigen Skala (Grundkenntnisse, gut und sehr gut) als sehr gut ein, 29% immerhin noch als gut, während 3% ihre Deutschkenntnisse nur als Grundkenntnisse einstufen. Nach dem exakten Fisher-Test ergeben sich sowohl

zwischen den Studien ($p = .003$, $V = 0.18$), als auch zwischen den beiden Standorten ($p = .032$, $V = 0.14$) signifikante Unterschiede mit kleinem Effekt. Dabei schätzen sowohl die Probanden der Studie ALSTER ihre Sprachkenntnisse besser ein (77% bzw. 63%) als auch die Probanden an der RUB (74% bzw. 67%). Bei FUNDAMENT wurden bzgl. der Sprache noch weitere Variablen erhoben. 80% der nicht-Muttersprachler, sprechen in der Familie überwiegend eine andere Sprache als Deutsch, im Freundeskreis wird dagegen bei 80% überwiegend Deutsch gesprochen.

Die Bildungsherkunft der Probanden wurde nach dem Konzept der Bildungsherkunft des DZHW (Middendorff et al., 2017) ermittelt. Dabei wird in vier Abstufungen die berufliche Bildung von Vater und Mutter geclustert:

- niedrig
  ein Elternteil hat einen beruflichen – nicht akademischen – Abschluss
- mittel
  beide Eltern haben einen beruflichen – nicht akademischen – Abschluss
- gehoben
  ein Elternteil hat einen akademischen Abschluss
- hoch
  beide Eltern haben einen akademischen Abschluss (Middendorff et al., 2017).

Im Gesamtdatensatz ergibt sich eine Verteilung der Bildungsherkunft nach dem genannten Schema von 20/39/20/20%[6] (niedrig/mittel/gehoben/hoch). Die Verteilung der 22. Sozialerhebung des DZHW (Kroher et al., 2023) für die Ingenieurwissenschaften (ohne Maschinenbau und Verfahrens-, Elektro- Informationstechnik) lautet: 10/34/30/26% (niedrig/mittel/gehoben/hoch). Daraus lässt sich schließen, dass die Probanden insgesamt einen niedrigeren Bildungshintergrund haben als in den Ingenieurwissenschaften üblich.

[6]Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

Damit zeichnet sich ein für das Ruhrgebiet typisches Bild einer heterogenen Studierendenschaft ab. Knapp die Hälfte der Probanden sind Nicht-Muttersprachler, von denen wiederum ein Drittel ihre Deutschkenntnisse nur als gut oder gar noch schlechter (3%) einschätzt. Dies lässt vermuten, dass es bei diesen Studierenden zu sprachlichen Problemen während des Studiums kommen kann. 60% der Probanden haben zudem eine niedrige oder mittlere Bildungsherkunft (nicht-akademisch, first-generation students), dieser Anteil liegt deutlich über dem für die Ingenieurwissenschaften üblichen Wert (44%). Die Bildungsherkunft lässt nicht nur sprachliche Schwierigkeiten erwarten, sondern insbesondere auch Herausforderungen bei der akademischen Enkulturation.

Die Gesamtstichprobe lässt sich hinsichtlich der Hochschulzugangsberechtigung (HZB) nach verschiedenen Aspekten beschreiben. Die durchschnittliche Abschlussnote der Stichprobe beträgt 2.7 ($SD = 0.6$) (Tab. 5.6) und zeigt keine signifikanten Unterschiede zwischen den beiden Studien. Allerdings zeigt der t-Test für beide Standorte ($95\% - CI[0.06, 0.31]$), $t(200) = 2.96$, $p = .003$, $d = 0.3$), dass die Probanden der RUB ($M = 2.5$, $SD = 0.6$) signifikant bessere Abschlussnoten haben als die Probanden der UDE ($M = 2.7$, $SD = 0.7$). In Tab. 5.6 sind auch die zuletzt erreichten Punkte in der Sekundarstufe II für die Fächer mit einer inhaltlichen Nähe zu den Ingenieurwissenschaften dargestellt. Es liegen nur Daten für FUNDAMENT vor, da diese Selbstabfrage in ALSTER nicht erfolgte. Im Durchschnitt erreichten die Probanden in den verschiedenen Fächern zwischen 8.8 und 10 Punkten.

Die Verteilung der Kurswahl (Grundkurs [GK] oder Leistungskurs [LK]) in der Oberstufe ist in Tab. 5.7 dargestellt, es sind keine signifikanten Unterschiede zwischen den Studien oder den Standorten zu erkennen. Die Kurswahl in Mathematik ist relativ ausgeglichen, als Pflichtfach musste eine der beiden Kursvarianten in der Sekundarstufe II gewählt werden. In Physik und Chemie wählten jeweils knapp 33% einen Leistungskurs, während 56% bzw. 62% keine entsprechenden Kurse in der Sekundarstufe II hatten. In den Fächern Informatik und Technik sind die Zahlen noch dramatischer, hier haben 80% bzw. 90% der Probanden in der Sekundarstufe II keinen Unterricht in diesen Fächern, während 18% bzw. 6% einen Leistungskurs besucht haben. Generell fällt auf,

Tabelle 5.6
Stichprobe - Note der Hochschulzugangsberechtigung und die zuletzt erreichten Fachpunktzahlen

	n	M	SD	Min	Max
HZB Note	534	2.7	0.6	1	4
Punkte Mathematik[a]	253	8.8	2.8	2	15
Punkte Physik[a]	107	9.6	2.9	2	15
Punkte Chemie[a]	96	9.7	2.9	1	15
Punkte Informatik[a]	45	10.0	2.7	5	15
Punkte Technik[a]	18	9.6	2.4	5	13

Anmerkungen. n = Stichprobengröße; M = Mittelwert; SD = Standardabweichung; Min = Minimum; Max = Maximum; HZB-Note = Note der Hochschulzugangsberechtigung.
[a] Die Daten stammen ausschließlich aus FUNDAMENT, in ALSTER wurden diese Daten nicht erhoben.

Tabelle 5.7
Stichprobe - Kurswahl verschiedener Fächer in der Oberstufe

	n	LK	GK	nicht belegt
Mathematik	503	235	268	
Physik	500	168	50	282
Chemie	492	167	20	305
Informatik	494	93	5	396
Technik	498	32	13	453

Anmerkungen. n = Stichprobengröße; LK = Leistungskurs; GK = Grundkurs.

dass – sofern eines der Fächer in der Oberstufe gewählt wurde – ein Leistungskurs in der Regel einem Grundkurs vorgezogen wurde.

Abschließend kann für die Hochschulzugangsberechtigung noch untersucht werden, auf welche Weise und in welchem Bundesland bzw. Land sie erworben wurde. 58% haben ihre Hochschulzugangsberechtigung durch ein Abitur an einem Gymnasium erworben, davon 54% in der Form G8. Die Art des Abiturs am Gymnasium unterscheidet sich zwischen den beiden Studien signifikant mit einem kleinen Effekt ($\chi^2(1) = 5.12$, $p = .024$, $\varphi = 0.12$): Während bei ALSTER 82% ein G8-Abitur angeben, sind es bei FUNDAMENT bereits 91%. Dieser

Unterschied lässt sich mit der Einführung von G8 und der zeitlichen Nähe des ersten G8-Abiturjahrgangs 2013 zur Studie ALSTER erklären. Weitere 28% der Probanden haben ihre Hochschulzugangsberechtigung an einer Gesamtschule erlangt, die übrigen Probanden haben ihre Hochschulreife über sonstige Wege (7%), ein Berufskolleg (5%) oder über den 2. Bildungsweg (2%) erworben. Der exakte Fisher-Test zeigt sowohl zwischen den Studien ($p < .001$, $V = 0.19$) als auch zwischen den beiden Standorten ($p = .025$, $V = 0.1$) signifikante Unterschiede mit kleinem Effekt. Beide Effekte sind darauf zurückzuführen, dass bei ALSTER und an der RUB die Probanden häufiger ein Gymnasium besucht haben (ca. 12% mehr), während bei FUNDAMENT und an der UDE im Vergleich häufiger die Gesamtschule besucht wurde (4% bis 6% mehr) und der sonstige Erwerb der Hochschulzugangsberechtigung häufiger war (6% bis 9% mehr). Die Probanden wurden auch gefragt, in welchem Bundesland sie ihre Hochschulzugangsberechtigung erworben haben. In der Stichprobe gaben 88% der Probanden NRW an, auf die anderen Bundesländer entfallen zusammen knapp 2%, die zweitgrößte Gruppe bilden diejenigen, die ihre Hochschulzugangsberechtigung im Ausland erworben haben (10%). Signifikante Unterschiede mit einem kleinen Effekt liegen nach dem exakten Fisher-Test zwischen den beiden Studien vor ($p = .01$, $V = 0.11$). Der Unterschied besteht darin, dass in ALSTER nur 5% der Probanden ihre Hochschulzugangsberechtigung im Ausland erworben haben, während es in FUNDAMENT 12% sind. Dieser Anstieg könnte mit dem starken Zuwachs der Asylanträge seit 2014 (Bundesamt für Migration und Flüchtlinge [BAMF], 2023) zusammenhängen und sich zeitversetzt in den Daten niederschlagen. Dies würde auch den deutlichen Anstieg der Nicht-Muttersprachler zwischen ALSTER (42%) und FUNDAMENT (52%) erklären. Für FUNDAMENT kann die Gruppe derjenigen, die ihre Hochschulzugangsberechtigung im Ausland erworben haben, genauer beschrieben werden. 39 Probanden haben dazu Angaben gemacht, um welches Land es sich handelt: Während die meisten Länder nur ein- oder höchstens zweimal vertreten sind (z.B. Türkei, Italien, Belgien oder Russland), stellt Syrien mit 26 Probanden (67%) die größte Gruppe. Dies deckt sich auch mit den bereits erwähnten hohen Zahlen von Asylsuchenden in Deutschland, bei denen Asylsuchende aus Syrien

seit 2014 jedes Jahr die größte Gruppe bilden (BAMF, 2023).

9% der Probanden haben vor dem Studium eine Berufsausbildung absolviert, davon 69% mit inhaltlicher Nähe zum Bauingenieurstudium. Vor Beginn des Bauingenieurstudiums haben 14% bereits Studienerfahrung gesammelt, davon haben 12% das Studium bereits abgeschlossen. In ALSTER wurde auch erhoben, ob es sich bei dem vorherigen Studium um ein Ingenieurstudium handelte. Von den 25 Probanden, die bereits im Vorfeld Studienerfahrung sammelten, war dies in 52% der Fälle ein anderes ingenieurwissenschaftliches Studium. 4 Probanden haben ihr vorheriges Studium abgeschlossen, die Hälfte davon im Bereich der Ingenieurwissenschaften.

Die Probanden wurden auch gefragt, ob sie vor Beginn ihres Studiums einen Vorkurs besucht haben, 44% der Probanden bejahten diese Frage. Es gibt einen signifikanten Unterschied mit einem kleinen Effekt zwischen der Nutzung der Vorkurse in den beiden Studien ($\chi^2(1) = 14.6$, $p < .001$, $\varphi = 0.16$), während bei FUNDAMENT nur 38% einen Vorkurs besucht haben, sind es bei ALSTER 54%. Auch zwischen den beiden Standorten gibt es einen signifikanten Unterschied, ebenfalls mit kleinem Effekt ($\chi^2(1) = 30.2$, $p < .001$, $\varphi = 0.23$), an der UDE haben 38% einen Vorkurs genutzt, an der RUB 66%. Die Ursache für diese unterschiedliche Nutzung kann anhand der Daten aus der Stichprobe nicht geklärt werden. In FUNDAMENT wurde nach der Art des Vorkurses gefragt: Von den 38%, die einen Vorkurs besucht haben, haben 21% den FUNDAMENT OV, 11% einen anderen OV und 69%[7] einen klassischen Präsenzvorkurs absolviert.

Zusammenfassend lässt sich sagen, dass es signifikante Unterschiede zwischen den beiden Studien und auch zwischen den beiden Standorten hinsichtlich verschiedener demographischer Variablen gibt. Da diese jedoch alle nur geringe Effekte aufweisen und als typische Jahrgangs- bzw. Kohortenunterschiede aufgefasst werden dürfen, können im Folgenden – sofern nicht eine getrennte Betrachtung der beiden Studien oder der Standorte erforderlich ist – die Daten aller Probanden als Gesamtdatensatz analysiert werden.

[7] Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

Kapitel 6

Ergebnisse und Interpretation

In diesem Kapitel werden die Ergebnisse der Untersuchung vorgestellt und interpretiert. Zunächst wird die Aufbereitung der erhobenen Daten erläutert, anschließend wird der Datensatz auf Ausreißer überprüft und fehlende Daten werden mit einem geeigneten Verfahren imputiert. Der aufbereitete Datensatz wird dann IRT skaliert. Dabei ist zu prüfen, ob der Datensatz überhaupt als gemeinsamer Datensatz skaliert werden kann oder ob eine Trennung nach Studien (FUNDAMENT und ALSTER) erforderlich ist. Anschließend werden die Fit-Werte der Items auf Modellpassung geprüft, erst dann ist eine IRT-Skalierung möglich. Verschiedene IRT-Modelle (1pl, 2pl, 3pl, 1- oder 2-/3-dimensional) werden skaliert und auf Modellpassung geprüft. Das Modell mit der besten Passung zum vorliegenden Datensatz wird für die Niveaumodellierung herangezogen. Im Anschluss an die Niveaumodellierung werden Unterschiede zwischen den einzelnen MZP identifiziert und interpretiert. Das Kapitel schließt mit der Beantwortung der in Kap. 3 postulierten Forschungsfragen.

6.1 Aufbereitung und Kodierung der Daten

Die Testhefte wurden mit Hilfe der Software ›TeleForm‹ (Electric Paper Informationssysteme GmbH, o. J.) erstellt, von den Probanden bearbeitet und anschließend digitalisiert. Sowohl die handschriftlichen Ein-

tragungen (z.B. Geburtsjahr, Abiturnote oder Punkte in den einzelnen Schulfächern) als auch die gesetzten Kreuze bei den Multiple-Choice-Aufgaben werden von der Software erkannt. Anschließend findet die Speicherung der Rohdaten in einer ›SPSS‹-Datei (IBM, 2022) statt. Aus ›SPSS‹ werden die Daten als .csv-Datei mit kommagetrennten Werten (engl.: *comma separated values*) exportiert und anschließend in ›Microsoft EXCEL‹ weiterverarbeitet. Alle Analysen werden mit der Statistiksoftware ›R‹ (R Core Team, 2023) oder ›SPSS‹ (IBM, 2022) durchgeführt. Als erster Schritt der Weiterverarbeitung werden die gesamten Rohdaten überprüft und gegebenenfalls korrigiert.

Fehlende Werte (engl.: *missing data/values*) werden aus dem englischen als *not available* (NA) kodiert. Von fehlenden Werten oder auch *Nonresponses* wird gesprochen, wenn ein Proband an einer Datenerhebung teilnimmt, aber nicht alle Fragen substanziell beantwortet (Bergmann & Franzese, 2020). Dies gilt sowohl für ausgelassene Antworten der Probanden als auch für nicht lesbare oder mehrfach angekreuzte Items. In Anlehnung an weitere Forschungsprojekte an der UDE (u.a. ALSTER und *Chemie, Sozialwissenschaften und Ingenieurwissenschaften: Studienerfolg und Studienabbruch* (CASSIS)) werden die ersten drei fehlenden Werte am Ende des Testheftes als ›0‹ und damit als falsch kodiert, weitere nachfolgende fehlende Werte behalten die Kodierung ›NA‹. Da eine gewissenhafte Bearbeitung der einzelnen Testhefte bei einer großen Anzahl von fehlenden Werten bezweifelt werden muss, werden Probanden mit mehr als 30% fehlenden Werten – in einem gesamten Testheft an einem MZP – aus dem Datensatz entfernt. Auch hier handelt es sich um heuristische Evidenz aus den genannten Forschungsprojekten der UDE.

Alle zur Kompetenzdiagnostik eingesetzten Items (sowohl der Technischen Mechanik als auch der Rechenfähigkeit) werden im Rahmen dieser Untersuchung dichotom kodiert. Die in der Studie FUNDAMENT verwendeten Items weisen alle ein geschlossenes Antwortformat auf. Dabei ist zu unterscheiden zwischen klassischen Multiple-Choice-Aufgaben mit nur einer richtigen Antwort und Multiple-Choice-Aufgaben mit mehreren richtigen Antwortalternativen (Moosbrugger & Brandt, 2020). Items mit einer richtigen Antwort werden mit ›1‹ kodiert, während

falsche Antworten mit ›0‹ kodiert werden. Ähnlich wird bei Multiple-Choice-Aufgaben mit mehreren richtigen Antworten verfahren: Sofern alle richtigen Antwortalternativen ausgewählt wurden, wird entsprechend eine ›1‹ kodiert. Wurden hingegen eine oder mehrere richtige Antwortalternativen nicht bzw. eine oder mehrere falsche Antworten ausgewählt, so gilt das Item als nicht richtig gelöst und wird mit ›0‹ kodiert. Multiple-Choice-Aufgaben mit mehreren richtigen Antwortalternativen wurden nur am vierten MZP im Teilbereich TM_4 (*TM4_10, TM4_18, TM4_19, TM4_20*) verwendet. Das Item *TM4_21* besteht aus drei Teilaufgaben mit jeweils einer klassischen Multiple-Choice-Aufgabe. Bei richtiger Beantwortung aller drei Teilaufgaben wird dieses Item mit ›1‹ kodiert, bei einer oder mehreren falschen Teilaufgaben mit ›0‹.

Auch in der Studie ALSTER wird überwiegend ein geschlossenes Antwortformat mit klassischen Multiple-Choice-Aufgaben eingesetzt, bei den Rechenfähigkeits-Items wird jedoch ein offenes Antwortformat verwendet. Nach J. Müller et al. (2018) werden diese Items mit Hilfe eines Kodierhandbuchs dichotom kodiert.

Im Anschluss an die Kodierung werden zwei weitere Schritte der Datenaufbereitung durchgeführt. Zunächst erfolgt eine Überprüfung des Datensatzes auf auffällige Werte (Ausreißer/Extremwerte), die gegebenenfalls aus dem Datensatz entfernt werden. Der so bereinigte Datensatz wird anschließend imputiert, um zufällig auftretende fehlende Werte zu eliminieren.

6.1.1 Extremwertanalyse

Eine Extremwertanalyse soll dazu dienen, auffällige Werte (Ausreißer/ Extremwerte) im Datensatz zu identifizieren, die ggf. aus dem Datensatz entfernt werden sollten. Dies geschieht auf der Basis der in den jeweiligen Teilbereichen erreichten Summenscores (engl. *total score* [TS]) an den einzelnen MZP. Unter Summenscores wird in dieser Arbeit die zeilenweise Aufsummierung der dichotomen Itemwerte zu einem Summenwert, auch Testwert genannt (Moosbrugger et al., 2020), verstanden.

Ein visuelles Verfahren zur Veranschaulichung von Ausreißern bzw. Extremwerten ist das sogenannte Box-Whisker-Diagramm nach Tukey (1977) (Bortz & Schuster, 2010; Eid et al., 2017; Sedlmeier & Burkhardt,

2021). Diese Darstellungsweise ist relativ robust gegenüber Ausreißern, da die Kennwerte für die Anfertigung (bspw. Median oder die Quartile) durch eine geringe Anzahl von Ausreißern nicht oder nur wenig beeinflusst werden (Bortz & Schuster, 2010).

Kernelement des Box-Whisker-Diagramms ist der Kasten (engl.: *box*), dessen Kastenende unten das erste Quartil (Q_1) und oben das dritte Quartil (Q_3) beschreibt (Eid et al., 2017). Die Differenz zwischen diesen beiden Quartilen gibt die Höhe des Kastens an und wird als Interquartilabstand (IQA) bezeichnet (Eid et al., 2017):

$$IQA = Q_3 - Q_1 \tag{6.1}$$

Die Endpunkte des IQA werden Scharniere (engl.: *hinges*) genannt (Sedlmeier & Burkhardt, 2021). Der Median wird im Kasten durch eine waagerechte Linie markiert (Eid et al., 2017). Aus seiner Lage lässt sich ableiten, ob es sich um eine symmetrische oder asymmetrische Verteilung handelt:

- symmetrisch: Median in der Mitte des Kastens
- linkssteile Verteilung: Median näher am unteren Kastenende
- rechtssteile Verteilung: Median näher am oberen Kastenende (Eid et al., 2017).

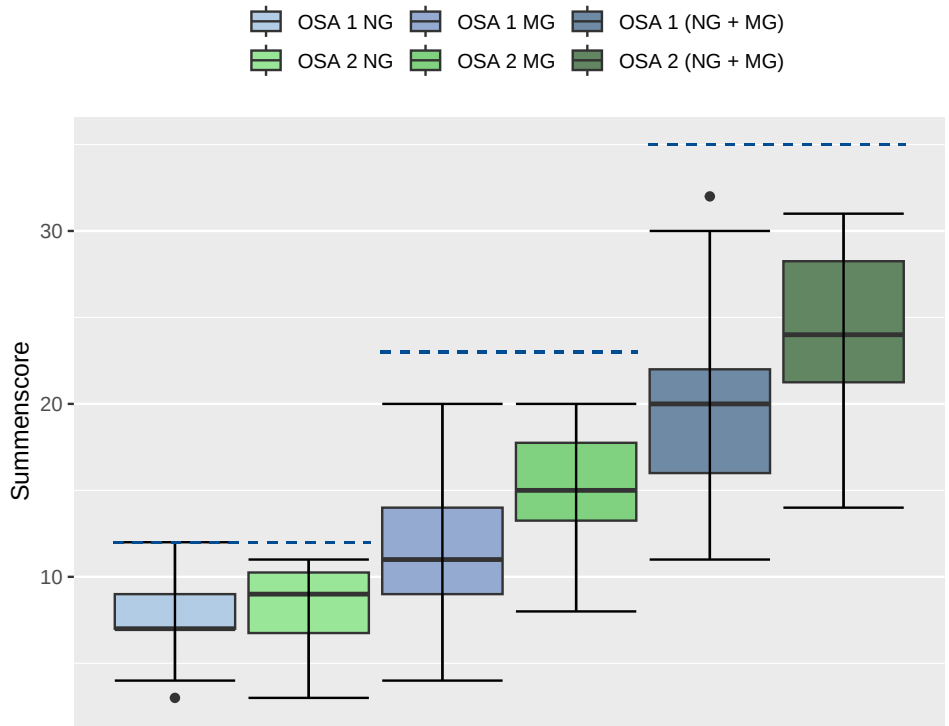
Die von den beiden Kastenenden ausgehenden Linien, auch Schnurrbarthaare (engl.: *whiskers*) genannt, sind ebenfalls zentrale Elemente des Box-Whisker-Diagramms. Sie können als weiteres Streuungsmaß verstanden werden, die durch imaginäre Zäune (engl.: *fences*) begrenzt sind (Sedlmeier & Burkhardt, 2021). Die maximale Länge der Linie kann berechnet werden. Die Linie am oberen Kastenende darf den Wert $Q_3 + 1,5 \cdot IQA$ nicht überschreiten (Eid et al., 2017). Wenn dieser Wert im Datensatz vorkommt, endet die Linie genau an dieser Stelle, wenn nicht, endet sie beim nächstkleineren Wert (Eid et al., 2017). Analog wird mit der Linie am unteren Kastenende verfahren. An dieser Stelle ist der Wert mindestens $Q_1 - 1,5 \cdot IQA$, wenn dieser Wert im Datensatz nicht vorhanden ist, beginnt die Linie beim nächsthöheren

Wert (Eid et al., 2017). Bortz und Schuster (2010) sprechen in diesem Zusammenhang von der oberen bzw. unteren Ausreißergrenze.

Dementsprechend werden alle Werte, die nicht innerhalb der Zäune liegen, als Ausreißer betrachtet (Sedlmeier & Burkhardt, 2021). Ausreißer können somit definiert werden als Werte größer $Q_3 + 1,5 \cdot IQA$ bzw. kleiner $Q_1 - 1,5 \cdot IQA$ (Eid et al., 2017). Wenn Ausreißer besonders stark nach oben oder unten abweichen, spricht man von Extremwerten (Eid et al., 2017). Diese sind wiederum definiert als Werte größer $Q_3 + 3 \cdot IQA$ bzw. kleiner $Q_1 - 3 \cdot IQA$ (Eid et al., 2017). Ausreißer werden als Punkte außerhalb der Ausreißergrenze dargestellt. Gibt es mehrere Ausreißer mit dem gleichen Zahlenwert, bleibt die Darstellung gleich.

Abbildung 6.1

Box-Whisker-Diagramm der Summenscores an MZP 1



Anmerkungen. OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen.

In Abb. 6.1, 6.2, 6.3 und 6.4 sind die Box-Whisker-Diagramme der erreichten Summenscores für die MZP 1-4 dargestellt. Zur besseren Ver-

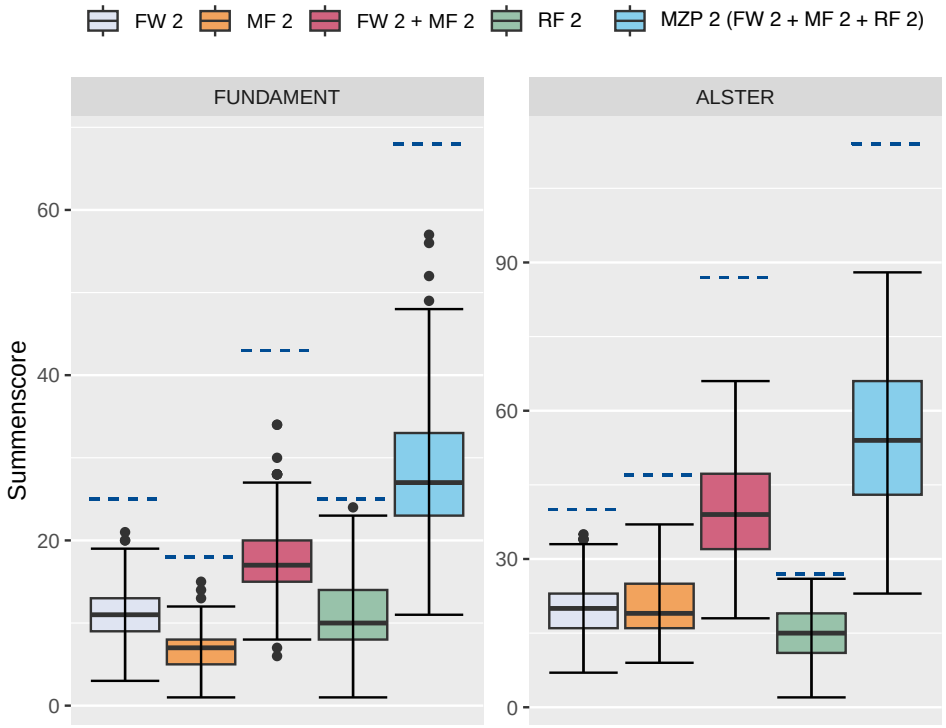
anschaulichung wurden in die Diagramme gestrichelte Linien eingefügt, die jeweils den maximal erreichbaren Summenscore der Teilbereiche an dem jeweiligen MZP darstellen. Bei allen auffälligen Werten handelt es sich um Ausreißer, keiner dieser Werte erfüllt die oben genannten Kriterien eines Extremwertes.

Die Abb. 6.1 stellt die erreichten Summenscore des OSA dar. Dabei ist der OSA 1 in blau, also vor Abschluss des OV, und der OSA 2 nach Absolvierung von Teilen oder des gesamten OV in grün dargestellt (siehe Kap. 2.3.2). Es ist zu beachten, dass vereinzelt Probanden nur einen Teil des OSA absolviert haben, beispielsweise nur den OSA 1 NG aber nicht den OSA 1 MG. Diese Probanden werden in der summierten Betrachtung nicht berücksichtigt ($OSA\ 1 = OSA\ 1\ NG + OSA\ 1\ MG$ bzw. $OSA\ 2 = OSA\ 2\ NG + OSA\ 2\ MG$). An MZP 1 sind zwei Ausreißer zu erkennen (siehe Abb. 6.1). Einer dieser beiden Ausreißer tritt im Inhaltsbereich OSA 1 NG auf und liegt im unteren Bereich der Skala (ein Proband). Der zweite Ausreißer befindet sich dagegen im oberen Teil der Skala bei der Gesamtbetrachtung von OSA 2.

Die in den Fachwissenstests an den MZP 2-4 erreichten Summenscores sind in den Abbildungen 6.2, 6.3 und 6.4 dargestellt. Auf der linken Seite sind jeweils die Ergebnisse der Studie FUNDAMENT, auf der rechten Seite die Ergebnisse der Studie ALSTER abgebildet. Zu beachten ist, dass die y-Achsen aufgrund der unterschiedlichen Anzahl der verwendeten Items (siehe 5.2) keine einheitliche Skalierung aufweisen. Analog zur Darstellung des Box-Whisker-Diagramms des OSA wurden auch in den Abbildungen der MZP 2-4 diejenigen Probanden nicht in die Darstellung der aggregierten Betrachtung einbezogen, die einen der Inhaltsbereiche nicht bearbeitet haben (bspw. Bearbeitung von FW 2 und MF 2, aber Auslassung von RF 2).

Die detaillierte Betrachtung von FUNDAMENT an MZP 2 zeigt, dass in allen Teilbereichen Ausreißer auftreten (siehe Abb. 6.2). Diese befinden sich überwiegend im oberen Bereich der Skala, lediglich in der summierten Betrachtung von FW 2 und MF 2 gibt es zwei Ausreißer nach unten (zwei Probanden). Bei dem gleichen MZP zeigen sich für ALSTER ausschließlich im Teilbereich FW 2 Ausreißer, diese liegen allesamt im oberen Bereich der Skala.

Abbildung 6.2
Box-Whisker-Diagramm der Summenscores an MZP 2



Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; MZP = Messzeitpunkt.

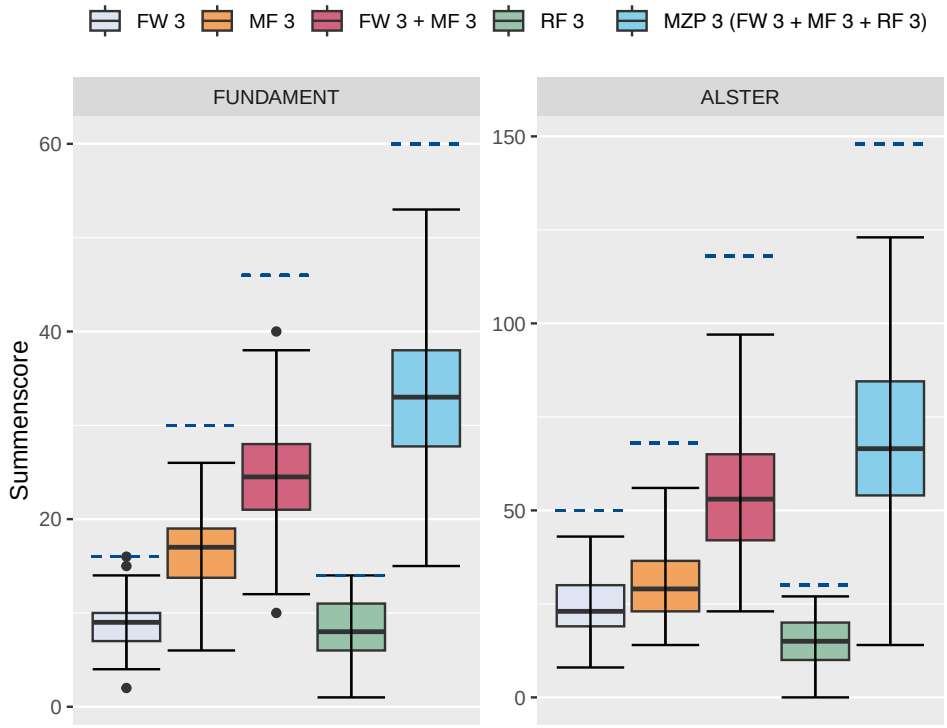
Abb. 6.3 zeigt, dass an MZP 3 nur in FUNDAMENT Ausreißer auftreten. Sowohl bei der Betrachtung des Teilbereichs FW 3 als auch bei der Summenbetrachtung von FW 3 und MF 3. Diese Ausreißer treten im unteren und oberen Bereich der Skala auf. Bei den Ausreißern im unteren Teil handelt es sich um den gleichen Probanden.

Abschließend zeigt sich ein Ausreißer in dem Teilbereich TM_4 an MZP 4 FUNDAMENT im oberen Teil der Skala (siehe Abb. 6.4).

Der Umgang mit den zuvor genannten Ausreißern soll nachfolgend diskutiert werden. Eid et al. (2017) empfehlen die Überprüfung der Werte auf eine sinnvolle Interpretierbarkeit:

> Hat man gute Gründe dafür anzunehmen, dass der Ausreißerwert dadurch zustande kommt, dass die Person nicht sorgfältig an der Untersuchung teilgenommen hat, so kann

Abbildung 6.3
Box-Whisker-Diagramm der Summenscores an MZP 3



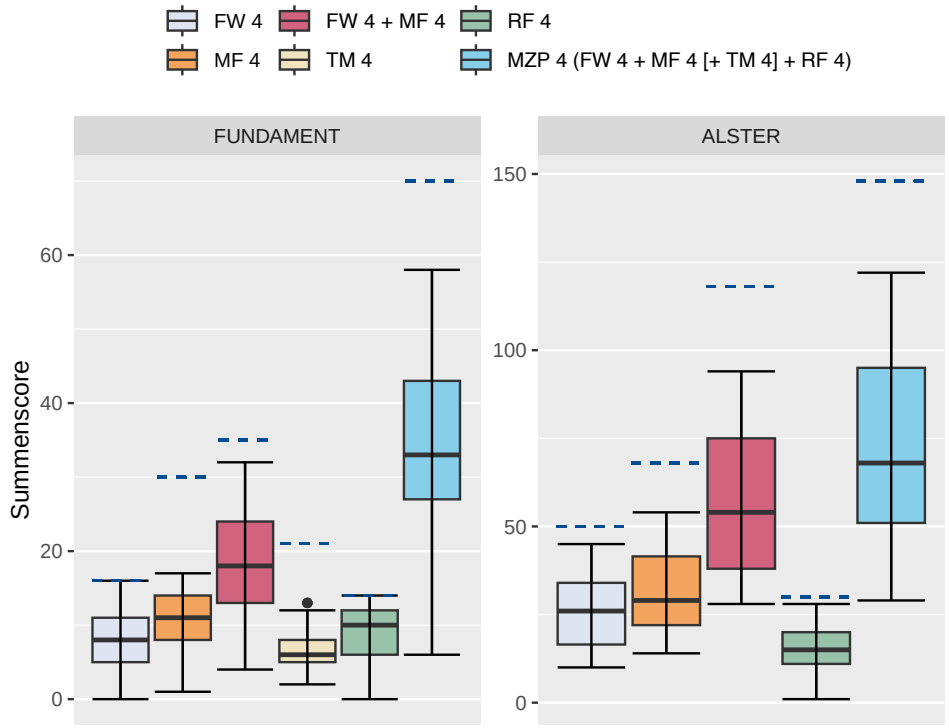
Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; MZP = Messzeitpunkt.

> der Wert aus dem Datensatz entfernt werden, da er keiner zuverlässigen Angabe entspricht. (Eid et al., 2017)

Bei erreichten Punktzahlen, die größer sind als $Q_3 + 1,5 \cdot IQA$ und damit oberhalb der Ausreißergrenze am oberen Ende der Skala liegen, kann nicht von mangelnder Sorgfalt bei der Bearbeitung der Testinstrumente gesprochen werden, auch wenn es sich nach der Analyse um Ausreißer handelt.

Anders verhält es sich mit den Summenscores, die als Ausreißer im unteren Bereich der Skala liegen und bei denen von einer mangelnden Sorgfalt bei der Bearbeitung ausgegangen werden könnte. Es zeigt sich jedoch, dass die Abweichungen von den Ausreißergrenzen minimal sind (1 oder 2 Punkte) und die Probanden auch in anderen Teilbereichen der jeweiligen Testhefte keine auffälligen Summenscores erzielt haben. Es

Abbildung 6.4
Box-Whisker-Diagramm der Summenscores an MZP 4



Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4; MZP = Messzeitpunkt.

wird daher angenommen, dass die Ausreißer nicht auf mangelnde Sorgfalt bzw. Motivation zurückzuführen sind. Vielmehr ist davon auszugehen, dass die Kompetenzen der Probanden nicht ausreichend waren, um einen höheren Summenscore zu erreichen.

Auffällig ist auch, dass an den MZP 2-4 einige der unteren Ausreißergrenzen im Summenscorebereich von null, eins oder zwei liegen (siehe Abbildungen 6.2, 6.3 und 6.4). Diese Lösungen stellen nach der durchgeführten Analyse keine Ausreißer dar und liegen somit im Bereich der akzeptablen Streuung. Dennoch sind derart niedrige Summenscores in einzelnen Teilbereichen auffällig und lassen Zweifel an der Sorgfalt der Bearbeitung aufkommen. Allerdings zeigt sich auch hier, dass die betroffenen Probanden in anderen Teilbereichen Items erfolgreich gelöst haben. Entsprechend wird auch hier angenommen, dass die niedrigen

Ergebnisse auf mangelnde Kompetenzen der Probanden zurückzuführen sind.

Tabelle 6.1
Stichprobengröße

	$n_{Probanden}$	n_{Items}
FUNDAMENT		
MZP 1		
OSA 1		
NG	33	12
MG	31	23
OSA 2		
NG	16	12
MG	14	23
MZP 2		
FW/MF	296	43
RF	274	25
MZP 3		
FW/MF	116	45
RF	114	14
MZP 4		
FW/MF/TM_4	105	56
RF	105	14
ALSTER		
MZP 2		
FW/MF	188	87
RF	193	27
MZP 3		
FW/MF	183	118
RF	184	30
MZP 4		
FW/MF	75	118
RF	77	30

Anmerkungen. $n_{Probanden}$ = Anzahl der Probanden; n_{Items} = Anzahl der Items; MZP = Messzeitpunkt; OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4.

Insgesamt werden nach der erfolgten Extremwertanalyse keine Probanden aus dem Datensatz entfernt. Die jeweilige Anzahl der Probanden ($n_{Probanden}$) sowie der eingesetzten Items (n_{Items}) sind in Tab. 6.1 nach Studie, MZP und Teilbereich aufgelistet.

6.1.2 Imputation

Aus drei Gründen ist es problematisch, wenn in einem Datensatz fehlende Werte vorliegen:

1. die Reliabilität (Genauigkeit für die Schätzung der Population) sinkt
2. die Schätzung kann systematisch verzerrt sein
3. bestimmte statistische Verfahren sind auf große Stichproben angewiesen und Signifikanztests verlieren mit abnehmender Probandenanzahl an Teststärke (Burkhardt et al., 2022)

Deswegen muss diskutiert werden, wie mit fehlenden Werten im Datensatz verfahren wird. Burkhardt et al. (2022) empfehlen eine nachträgliche Erfassung der fehlenden Werte oder eine Erhebung zusätzlicher Fälle. Dies ist jedoch im Rahmen dieser Arbeit aus organisatorischen und inhaltlichen Gründen nicht möglich, u.a. könnte ein erneutes Vorlegen der Items zu einer Verfälschungen der Ergebnisse führen (Burkhardt et al., 2022). Daher wird das Verfahren der Imputation in Betracht gezogen, um die fehlenden Werte aus dem Datensatz zu entfernen.

In der Methodenlehre werden drei Arten von fehlenden Werten unterschieden:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR) (Burkhardt et al., 2022; Eid et al., 2017).

Für MCAR bestehen keine Abhängigkeiten, d.h. weder die betrachtete Variable noch andere Variablen haben einen Einfluss auf das Vorhandensein eines fehlenden Wertes (Bergmann & Franzese, 2020; Eid et al., 2017). Das Fehlen ist somit rein zufällig, statistische Kennwerte können geschätzt werden, allerdings beeinflusst diese Art fehlender Werte die Reliabilität und Teststärke, da die Fallzahlen beeinflusst werden (Burkhardt et al., 2022). Die MCAR-Annahme ist strikter als MAR oder MNAR.

Wenn die fehlenden Werte von einer anderen Variablen abhängen, aber die Ausprägung innerhalb der Variablen selbst zufällig ist, spricht man von MAR (Bergmann & Franzese, 2020; Burkhardt et al., 2022). Wie Bergmann und Franzese (2020) hervorheben, ist die Begrifflichkeit ›zufällig‹ in diesem Zusammenhang irreführend, zutreffender wäre die Bezeichnung ›bedingt zufällig‹. Bei dieser Art von fehlenden Werten ist eine Verzerrung der Schlussfolgerungen für die Gesamtpopulation möglich (Burkhardt et al., 2022).

Von MNAR spricht man, wenn die fehlenden Werte systematisch auftreten (Burkhardt et al., 2022). Es besteht eine Abhängigkeit der fehlenden Werte von der Variablen selbst, die nicht durch die Ausprägung anderer Variablen erklärt werden kann (Bergmann & Franzese, 2020; Eid et al., 2017). Eine Unterscheidung zwischen MAR und MNAR ist in der Regel nicht eindeutig möglich (Wirtz, 2004). Für eine empirische Bestimmung müssten die Ausprägungen der fehlenden Werte bekannt sein, daher erfolgt die Zuordnung auf inhaltlicher Basis des Forschungsgegenstandes (Wirtz, 2004). Je mehr erklärende Variablen für die fehlenden Werte berücksichtigt werden können, desto größer ist jedoch die Wahrscheinlichkeit, dass es sich um MAR handelt (Bergmann & Franzese, 2020).

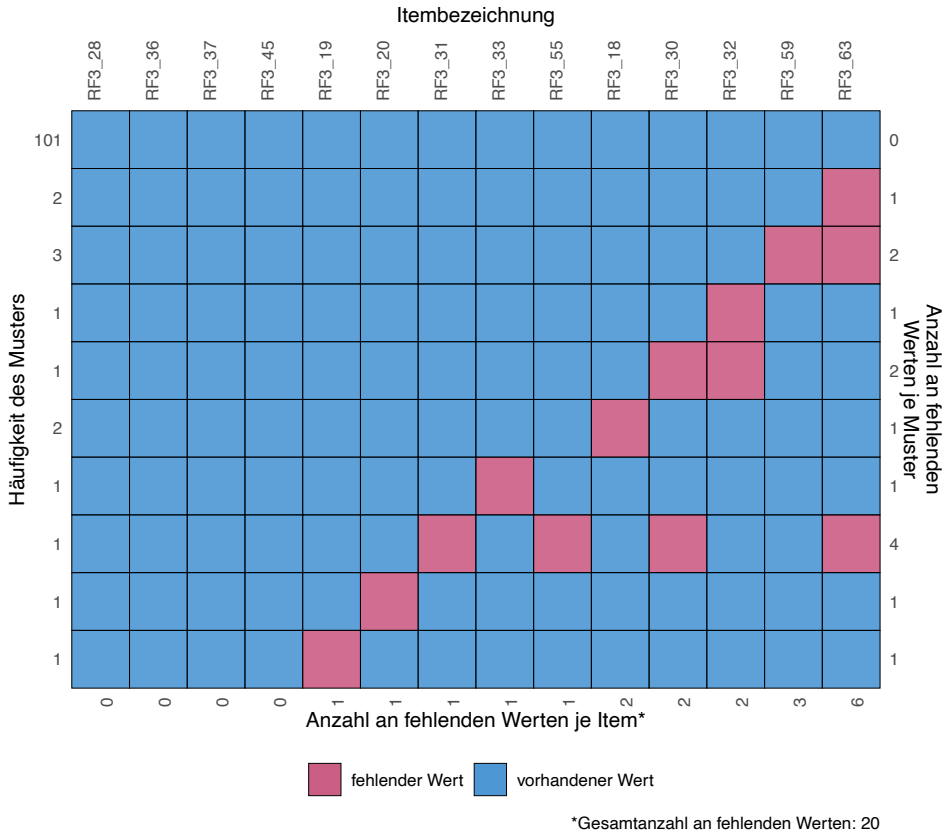
In einem ersten Schritt werden Umfang und Art der fehlenden Werte im Datensatz bestimmt. Da in den Studien FUNDAMENT und ALSTER eine unterschiedliche Anzahl an Items eingesetzt wurde (siehe Tab. 6.1), werden die Studien hinsichtlich der Diagnose der fehlenden Werte und der anschließenden Imputation getrennt betrachtet. Gleiches gilt für die MZP, die nochmals in die Inhaltsbereiche Technische Mechanik (Teilbereiche: FW, MF und TM_4) und Mathematik (Re-

chenfähigkeit) unterteilt werden. In der Studie ALSTER liegen für die Rechenfähigkeit-Items keine fehlenden Werte vor. Bei MZP 1 erfolgt ebenfalls eine Trennung in die einzelnen OSA (OSA 1 und OSA 2) sowie in die Inhaltsbereiche naturwissenschaftliche oder mathematische Grundlagen. Demographische Variablen werden bei dieser Analyse und der Imputation nicht berücksichtigt. Wie bereits in Kap. 6.1 erwähnt, wurden Probanden mit mehr als 30% fehlenden Werten an einem MZP bereits aus dem Datensatz entfernt und werden entsprechend nicht mehr berücksichtigt.

Die fehlenden Werte können in Abhängigkeit von den relevanten Items eines MZP in Form einer Matrix angegeben werden. Zur besseren Veranschaulichung wird diese Matrix im Folgenden exemplarisch für den dritten MZP der Studie FUNDAMENT für die Rechenfähigkeit-Items grafisch als Musterdiagramm der fehlenden Werte dargestellt (siehe Abb. 6.5) und erläutert. Dieses Diagramm wurde mit dem R-Paket ›ggmice‹ (Oberman, 2023) erstellt, dessen Funktion ›plot_pattern‹ an die Anforderungen dieser Arbeit angepasst wurde. Die Spalten des Musterdiagramms der fehlenden Werte zeigen die verschiedenen Items, die Zeilen die verschiedenen Muster des Auftretens der fehlenden Werte. Ein fehlender Wert wird in diesem Diagramm rot dargestellt, ein vorhandener Wert blau. In der Matrix hingegen ist ein fehlender Wert mit ›1‹ und ein vorhandener Wert mit ›0‹ gekennzeichnet. Auf der unteren X-Achse wird die Anzahl der fehlenden Werte je Item aufgetragen, wobei die Items nach der Häufigkeit ihres Auftretens geordnet sind. Auf der rechten Y-Achse ist die Anzahl der fehlenden Werte je Muster, d.h. je Zeile, angegeben, während auf der linken Y-Achse die Häufigkeit dieser Muster abzulesen ist. In dem abgebildeten Beispiel trifft die erste Zeile, in der kein einziger fehlender Wert auftritt, auf insgesamt 101 Probanden zu. Während das dritte Muster mit zwei fehlenden Werten (Items: *RF3_59* und *RF3_63*) insgesamt dreimal im Datensatz vorkommt. Das hier gezeigte Diagramm zeigt ein unauffälliges Muster, d.h. es ist keine Systematik erkennbar. Alle MZP weisen in beiden Studien fehlende Werte auf, auch hier ist keine Systematik zu identifizieren.

Um zu überprüfen, um welche Art von fehlenden Werten es sich handelt, wird der Test von Little (1988) herangezogen. Dieser Test

Abbildung 6.5
Musterdiagramm der fehlenden Werte - FUNDAMENT MZP 3 RF



Anmerkungen. RF = Rechenfähigkeit.

prüft die Mittelwerte für jede Variable mit und ohne fehlende Werte auf signifikante Unterschiede (Little, 1988). Die betrachtete Nullhypothese (H_0) lautet in diesem Zusammenhang: Die Mittelwerte unterscheiden sich nicht (Little, 1988). Die Alternativhypothese (H_1) hingegen besagt, dass sich die entsprechenden Mittelwerte unterscheiden (Little, 1988). In diesem Fall wäre also die Nullhypothese widerlegt, die fehlenden Werte sind nicht MCAR, sondern MAR oder MNAR. Die Überprüfung erfolgt auf Basis des zuvor festgelegten Signifikanzniveaus (α) (Bortz & Schuster, 2010; Döring & Bortz, 2016; Eid et al., 2017). Übliche Werte sind $\alpha = .05$, $\alpha = .01$ oder $\alpha = .001$, je nachdem, wie problematisch die Folgen der Ablehnung der Alternativhypothese sind, wird ein möglichst niedriges α gewählt (Bortz & Schuster, 2010; Döring & Bortz, 2016).

Der vorliegende Test wird mit $\alpha = .05$ durchgeführt, die Berechnung erfolgt mit der Statistiksoftware ›SPSS‹.

Deskriptive Statistiken der fehlenden Werte und die Ergebnisse des MCAR-Tests nach Little (1988) finden sich in Tab. 6.2. Dort sind je nach Studie, MZP und Teilbereich die Anzahl der aufgetretenen fehlenden Werte (n_{NA}) und die relative Häufigkeit (engl. *frequency* [f]) für das Auftreten der fehlenden Werte (f_{NA}) angegeben. Ebenso wird die Anzahl der vollständigen Fälle (engl. *complete cases* [cc]) und die entsprechende relative Häufigkeit (f_{cc}) aufgeführt. Die Statistiken des MCAR-Tests nach Little (1988) umfassen die χ^2-Werte, die Anzahl der Freiheitsgrade (df) und den zugehörigen Irrtumswahrscheinlichkeiten (engl. *probability* [p]). Der Tab. 6.2 ist zu entnehmen, dass die Anzahl der fehlenden Werte mit der Anzahl der Probanden und der Anzahl der Items zunimmt (siehe Tab. 6.1). Die relative Häufigkeit des Auftretens fehlender Werte liegt im niedrigen einstelligen Bereich ($f_{NA} = 1$ bis 7%). Diese geringe relative Häufigkeit der fehlenden Werte reicht jedoch aus, um die Anzahl und damit die relative Häufigkeit der vollständigen Fälle zum Teil erheblich zu beeinflussen. Schließlich ist für die vollständigen Fälle die Verteilung der fehlenden Werte entscheidend. So beeinflussen einzelne Probanden mit mehreren fehlenden Werten die relative Häufigkeit der vollständigen Fälle weniger stark als einzelne Probanden mit bspw. nur einem fehlenden Wert. Nach Urban und Mayerl (2018) kann ein Datensatz ohne weitere Bearbeitung verwendet werden, wenn in weniger als 5% der Fälle fehlende Werte auftreten, also 95% vollständige Fälle vorliegen. Dieser Wert wird in keinem Teilbereich an keinem MZP erreicht, daher sollte der Datensatz optimiert werden. Lässt man den ersten MZP wegen der geringen Probandenzahl unberücksichtigt, so weist die Studie FUNDAMENT mindestens 74% vollständige Fälle auf, bei ALSTER liegen diese Werte deutlich niedriger ($f_{cc} = 36$ bis 69%). Eine Erklärung hierfür liegt in der großen Anzahl der eingesetzten Items in den ALSTER-Testungen (siehe Tab. 6.1). Bei den Ergebnissen des MCAR-Tests nach Little (1988) ist der p-Wert ausschlaggebend für die Beibehaltung oder Ablehnung von H_0. Keiner der aufgelisteten p-Werte liegt unterhalb des zuvor festgelegten Signifikanzniveaus von $\alpha = .05$, genauer gesagt es liegen keine signifikanten Ergebnisse vor. H_0 wird

dementsprechend nicht abgelehnt, d.h. es wird angenommen, dass die vorhandenen fehlenden Werte der Art MCAR entsprechen.

Tabelle 6.2
Deskriptive Statistiken der fehlenden Werte und Ergebnisse des MCAR-Tests nach Little (1988)

	n_{NA}	f_{NA}	n_{cc}	f_{cc}	χ^2	df	p
FUNDAMENT							
MZP1							
OSA 1							
NG	14	.04	29	.88	32.77	26	.169
MG	53	.07	18	.58	248.07	246	.451
OSA 2							
NG	3	.02	15	.94	15.00	9	.091
MG	13	.04	7	.50	77.37	127	.999
MZP 2							
FW/MF	343	.03	219	.74	2589.92	2590	.497
RF	178	.03	203	.74	1264.15	1226	.219
MZP 3							
FW/MF	76	.02	93	.80	796.27	827	.773
RF	20	.01	101	.89	118.94	112	.309
MZP 4							
FW/MF/TM_4	66	.01	80	.76	901.75	949	.862
RF	11	.01	98	.93	88.95	87	.422
ALSTER							
MZP 2							
FW/MF	1088	.07	68	.36	8728.27	8938	.942
MZP 3							
FW/MF	193	.01	102	.56	7327.89	7259	.282
MZP 4							
FW/MF	76	.01	52	.69	984.09	1936	.999

Anmerkungen. n_{NA} = Anzahl der fehlenden Werte; f_{NA} = rel. Häufigkeit der fehlenden Werte; n_{cc} = Anzahl der vollständige Fälle; f_{cc} = rel. Häufigkeit der vollständige Fälle; χ^2 = χ^2-Wert des MCAR-Tests nach Little (1988); df = Freiheitsgrade; p = Wahrscheinlichkeit; MZP = Messzeitpunkt; OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4.

Das Auftreten von fehlenden Werten vom Typ MCAR ist weniger

problematisch als die beiden anderen Varianten (Wirtz, 2004). Es gibt verschiedene Möglichkeiten, mit den völlig zufällig fehlenden Werten (MCAR) umzugehen. Die gängigste Möglichkeit ist ein sogenanntes Eliminationsverfahren, bei dem diejenigen Probanden aus dem Datensatz ausgeschlossen werden, für die fehlende Werte vorliegen (Wirtz, 2004). Dieser zeilenweise Ausschluss hat zur Folge, dass nur noch Probanden mit vollständigen Daten im Datensatz verbleiben (Wirtz, 2004). In der Folge reduziert sich die Anzahl der Probanden - je nach Häufigkeit der fehlenden Werte - zum Teil erheblich. Diese Problematik kann durch Imputationsverfahren ausgeschlossen werden, da durch die multiple Imputation die Analyse mit einem vollständigen Datensatz durchgeführt werden kann. Bei diesen Verfahren werden die fehlenden Werte durch möglichst plausible Werte ersetzt bzw. imputiert (Burkhardt et al., 2022; Wirtz, 2004). Eine detaillierte Beschreibung der multiplen Imputation geben Urban und Mayerl (2018). Im Rahmen dieser Arbeit wird mit dem *Predictive Mean Matching* (PMM) ein zufallsbasierter Imputationsalgorithmus verwendet. Dieser Algorithmus kann auch auf nominal- oder ordinalskalierte Daten angewendet werden (Burkhardt et al., 2022). Bei PMM handelt es sich um ein sogenanntes *hot-deck*-Verfahren, bei dem ein ähnlicher, aber vollständiger Fall im Datensatz gesucht wird, der als Spender für den Fall mit fehlenden Werten dient (Vink et al., 2014). Als Ähnlichkeitskriterium kann z.B. das Prinzip des Nächsten-Nachbarn (engl.: *nearest-neighbor principle*) verwendet werden, wobei aus den vollständigen Daten der Fall mit der besten Übereinstimmung zu einem Fall mit fehlenden Werten herangezogen wird (Vink et al., 2014). Falls mehrere ähnliche Fälle bezüglich dieses Kriteriums gefunden werden, wird ein Fall zufällig ausgewählt und der fehlende Wert auf dieser Basis imputiert (Burkhardt et al., 2022). Während der Imputation wird dieser Vorgang mehrmals wiederholt, sodass mehrere simulierte Teilstichproben entstehen. Aufgrund des *hot-deck*-Charakters des PMM-Algorithmus werden plausible Werte imputiert, die auch beobachtet wurden, d.h. bei dichotomen Items werden auch nur entsprechende Werte imputiert (Vink et al., 2014).

Die Ersetzung der fehlenden Werte erfolgt mit Hilfe des ›R‹-Pakets ›mice‹ (van Buuren & Groothuis-Oudshoorn, 2011) und der gleichnami-

gen Funktion. Bei dieser Funktion wird der PMM-Algorithmus (›method = "pmm"‹) ausgewählt und die Anzahl der simulierten Teilstichproben auf 20 (›m = 20‹) festgelegt. Grundsätzlich gilt: Je höher f_{NA}, desto höher sollte die Anzahl der simulierten Teilstichproben sein (Urban & Mayerl, 2018). Urban und Mayerl (2018) empfehlen 20 Durchläufe, da diese bereits zu robusten Ergebnissen führen. Aus diesen 20 simulierten Teilstichproben wird mit der Funktion ›complete‹, ebenfalls aus dem Paket ›mice‹, eine Teilstichprobe gezogen. Im Rahmen dieser Arbeit wird einheitlich die zehnte simulierte Teilstichprobe (›action = 10‹) verwendet und die entsprechenden Summenscores des imputierten Datensatzes ermittelt.

Da fehlende Werte definitionsgemäß unbekannt sind, erweisen sich auch Validierungsmöglichkeiten für Imputationsverfahren als problematisch und nicht vorhanden (Burkhardt et al., 2022). Dennoch empfehlen Burkhardt et al. (2022) eine Überprüfung der imputierten Werte durch eine grafische Darstellung, durch die weniger sinnvolle Werte identifiziert werden können. Da jedoch nur dichotome Variablen imputiert wurden und PMM als *hot-deck*-Methode bereits andere als die vorhandenen Werte ausschließt, wird auf die grafische Validierung verzichtet. Stattdessen wird der t-Test für abhängige Stichproben verwendet, um die Mittelwerte der Summenscores des Quelldatensatzes und des imputierten Datensatzes zu vergleichen. Aufgrund der Imputation dichotomer Werte ist zu erwarten, dass die Mittelwerte des imputierten Datensatzes größer sind als die des Quelldatensatzes, da bereits die Imputation einer einzigen richtigen Antwort (›1‹) zu einem größeren Mittelwert führt.

Von abhängigen Stichproben spricht man u.a., wenn gleiche Probanden zu verschiedenen Zeitpunkten untersucht werden (Messwiederholungen) (Bortz & Schuster, 2010; Burkhardt et al., 2022; Eid et al., 2017). Diese Ausgangssituation kann auf die zu untersuchenden Datensätze übertragen werden. Damit der t-Test für abhängige Stichproben angewendet werden kann, müssen zwei Voraussetzungen erfüllt sein:

1. voneinander unabhängige Messwerte innerhalb der Stichprobe (Eid et al., 2017)

2. Normalverteilung der Differenzwerte der beiden Gruppen (Bortz & Schuster, 2010; Eid et al., 2017)

Da die Probanden des Datensatzes zufällig aus der Grundgesamtheit gezogen wurden, ist die erste Bedingung erfüllt. Hinsichtlich der zweiten Annahme, der Normalverteilung, erweist sich der t-Test relativ robust gegenüber einer Verletzung dieser Annahme, sofern die Stichprobe ausreichend groß ist ($n > 30$) (Bortz & Schuster, 2010; Eid et al., 2017). Diese Annahme ist nur an MZP 1 OSA 1 MG und OSA 2 (NG und MG) nicht erfüllt. Zur Überprüfung der Normalverteilungsannahme für MZP 1 OSA 1 MG, sowie OSA 2 NG bzw. MG wird der Shaprio-Wilk-Test verwendet (Shapiro & Wilk, 1965). In allen drei Fällen ist der p-Wert nicht signifikant (bei einem Signifikanzniveau von $\alpha = .05$), sodass auch in diesem Fall von einer Normalverteilung ausgegangen werden kann. Die H_0 des t-Tests für abhängige Stichproben gibt an, dass die Mittelwerte der Summenscores des Quelldatensatzes und des imputierten Datensatzes identisch sind (Bortz & Schuster, 2010; Eid et al., 2017). Durch die Imputation der dichotomen Werte ist es naheliegend, dass auch die zu betrachtenden Summenscores größer werden. Somit lässt sich theoretisch begründen, dass die gerichtete H_1, die von größeren Mittelwerten des imputierten Datensatzes ausgeht, betrachtet wird (Bortz & Schuster, 2010; Eid et al., 2017). Auch hier wird eine Signifikanzgrenze von $\alpha = .05$ festgelegt. Signifikante Ergebnisse führen zur Ablehnung von H_0, d.h. die Mittelwerte des ursprünglichen Datensatzes sind kleiner als die mit den imputierten Werten. Cohen's d beschreibt beschreibt die Größe dieses Effekts bezogen auf die Gesamtpopulation (Cohen, 1988; Eid et al., 2017). Wie bereits in Kap. 5.5 erwähnt, wird d in drei Stärken unterteilt: kleiner Effekt ($|d| \geq 0.2$), mittlerer Effekt ($|d| \geq 0.5$) und großer Effekt ($|d| \geq 0.8$) (Cohen, 1988).

In Tab. 6.3 werden die Ergebnisse der Gegenüberstellung des Quelldatensatzes mit dem imputierten Datensatz getrennt nach Studien, MZP und Teilbereiche dargestellt. Der t-Test für abhängige Stichproben wird durch den t-Wert (t), die Freiheitsgrade (df) und die Wahrscheinlichkeit (p) beschrieben. Die zugehörige Effektstärke wird mit Cohen's d (d)

Tabelle 6.3
Ergebnisse des t-Tests für abhängige Stichproben und Mittelwerte der Summenscores des Quelldatensatzes und des imputierten Datensatzes

	t	df	p	d	M_Q	M_I	ΔM[a]
FUNDAMENT							
MZP 1							
OSA 1							
NG	1.79	32	.042	0.31	7.5	7.7	0.3
MG	3.27	30	.001	0.59	11.4	12.2	0.7
OSA 2							
NG	1.00	15	.167	0.25	8.3	8.4	0.1
MG	2.83	13	.007	0.76	14.9	15.5	0.6
MZP 2							
FW/MF	6.45	295	<.001	0.37	17.6	18.0	0.5
RF	6.55	273	<.001	0.40	11.4	11.7	0.3
MZP 3							
FW/MF	2.70	115	.004	0.25	24.8	25.1	0.3
RF	2.67	113	.005	0.25	8.2	8.4	0.2
MZP 4							
FW/MF/TM_4	3.03	104	.002	0.30	25.2	25.5	0.3
RF	1.52	104	.066	0.15	8.8	8.9	0.1
ALSTER							
MZP 2							
FW/MF	10.42	187	<.001	0.76	40.0	42.6	2.6
MZP 3							
FW/MF	5.98	182	<.001	0.44	54.7	55.2	0.5
MZP 4							
FW/MF	2.89	74	.003	0.33	57.5	58.1	0.6

Anmerkungen. t = t-Wert; df = Freiheitsgrade; p = Wahrscheinlichkeit; d = Cohen's d; M_Q = Mittelwert der Summenscores des Quelldatensatzes; M_I = Mittelwert der Summenscores des imputierten Datensatzes; $\Delta M = M_I - M_Q$; MZP = Messzeitpunkt; OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4.
[a] Abweichungen in den Differenzen sind auf Rundungsfehler zurückzuführen.

angegeben. Die Tabelle enthält auch die jeweiligen Mittelwerte des Summenscores des Quelldatensatzes M_Q und des imputierten Datensatzes M_I, sowie deren Differenz $\Delta M = M_I - M_Q$. Mit Ausnahme von FUNDAMENT MZP 1 OSA 2 NG und FUNDAMENT MZP 4 RF sind alle Ergebnisse signifikant. Bei den beiden letztgenannten kann das Ergebnis auf die hohe relative Häufigkeit vollständiger Fälle ($f_{cc} = 94\%$ bzw. 93%) und das geringe Auftreten von fehlenden Werten ($f_{NA} = 2\%$ bzw. 1%) zurückgeführt werden (siehe Tab. 6.2). Der Unterschied zwischen den Mittelwerten ist mit $\Delta M = 0.1$ Punkten gering. Bei den übrigen MZP und den entsprechenden Teilbereichen ist aufgrund der signifikanten p-Werte die Effektstärke d zu betrachten. Bei FUNDAMENT MZP 1 OSA 1 MG, FUNDAMENT MZP 1 OSA 2 MG und ALSTER MZP 2 zeigen sich mittlere Effekte ($d = 0.59$ bis 0.76), die übrigen sind als leichte Effekte ($d = 0.15$ bis 0.44) einzustufen. Diese leichten Effekte zeigen sich in einer Mittelwertdifferenz von 0.2 bis 0.6 Punkten, wobei die relative Häufigkeit vollständiger Fälle stark schwankt ($f_{cc} = 56$ bis 94%), während die relative Häufigkeit fehlender Werte zwischen 1% und 4% liegt. Bei FUNDAMENT MZP 1 OSA 1 MG und FUNDAMENT MZP 1 OSA 2 MG treten mittlere Effekte auf, die in einer Differenz der Mittelwerte zwischen 0.6 und 0.7 Punkten resultieren. In beiden Fällen gibt es nur wenige vollständige Fälle ($f_{cc} = 58\%$ bzw. 50%) und viele fehlende Werte (siehe Tab. 6.2), während f_{NA} bei 7% bzw. 4% liegt. Deutlich drastischer sind die Ergebnisse der Studie ALSTER an MZP 2. Wie bereits erwähnt, liegt auch hier ein mittlerer Effekt vor, der sich jedoch in einer Mittelwertdifferenz von 2.6 Punkten niederschlägt und damit deutlich über den anderen Differenzen liegt. Die Gründe hierfür sind wiederum in der Betrachtung der deskriptiven Statistik zu identifizieren (siehe Tab. 6.1 und 6.2). In absoluten Zahlen gibt es 1.088 fehlende Werte ($f_{cc} = 7\%$). Die zweitmeisten fehlenden Werte finden sich bei FUNDAMENT MZP 2, aber mit 343 fehlenden Werten ist die Anzahl nur ein Drittel so groß ($f_{NA} = 3\%$). Ein ähnliches Bild ergibt sich bei der Betrachtung der vollständigen Fälle. Während nur 36% vollständige Fälle vorliegen, sind es bei FUNDAMENT MZP 2 mehr als doppelt so viele ($f_{cc} = 74\%$). Diese beiden Umstände führen dazu, dass die Differenz der Mittelwerte erheblich von den anderen MZP

und Teilbereichen abweicht. Letztendlich ergibt sich erwartungsgemäß eine Mittelwertdifferenz aufgrund der Imputation. Diese liegt in einem akzeptablen Bereich von 0.1 bis 2.6 Punkten. Die Abweichung von 2.6 Punkten ist im Vergleich zwar deutlich größer, jedoch aufgrund der hohen Anzahl fehlender Werte in der Studie ALSTER am zweiten MZP durchaus akzeptabel.

Wie bereits der Quelldatensatz (siehe Kap. 6.1.1) soll auch der imputierte Datensatz auf auffällige Werte (Ausreißer/Extremwerte) untersucht werden. Hierzu werden erneut Box-Whisker-Diagramme als visuelles Verfahren zur Identifikation herangezogen, die entsprechenden Diagramme finden sich im Anhang A. Am ersten MZP (siehe Abb. A.1) gibt es drei Ausreißer, einen im oberen Bereich und zwei im unteren Bereich der Skala. Insgesamt zehn Ausreißer im oberen Bereich der Skala treten bei MZP 2 auf (12 Probanden) (siehe Abb. A.2). Vierzehn Ausreißer zeigen sich bei MZP 3. Fünf im oberen Bereich der Skala (acht Probanden) und fünf im unteren Bereich der Skala (sechs Probanden) (siehe Abb. A.3). Schließlich gibt es bei MZP 4 zwei Ausreißer im oberen Bereich der Skala (drei Probanden) (siehe Abb. A.4). Analog zu Kap. 6.1.1 verbleiben alle Probanden im Datensatz. Bei Ausreißern nach oben kann nicht von mangelnder Sorgfalt bei der Bearbeitung der Testinstrumente ausgegangen werden. Die verzeichneten Ausreißer im unteren Bereich der Skala liegen in den einzelnen Teilbereichen maximal zwei Punkte unter der Ausreißergrenze, sodass auch hier eher von mangelnder Kompetenz bei der Beantwortung der Fragen als von fehlender Sorgfalt bzw. Motivation ausgegangen werden kann. Somit bleibt auch nach der Imputation die Stichprobengröße erhalten (siehe Tab. 6.1).

6.2 IRT-Skalierung

In diesem Kapitel erfolgt die IRT-Skalierung des imputierten Datensatzes. Dazu wird zunächst überprüft, ob eine gemeinsame Betrachtung als Gesamtdatensatz zulässig ist. Diese Prüfung ist sowohl für die FUNDAMENT Haupterhebung und Nacherhebung als auch zwischen den beiden Studien FUNDAMENT und ALSTER vorzunehmen. An-

schließend werden die Fit-Werte (wMNSQ-Infit und wMNSQ-Outfit) der verwendeten Items überprüft. Items mit auffälligen Fit-Werten werden von der IRT-Skalierung ausgeschlossen. Bei der IRT-Skalierung werden verschiedene Modelle (1pl, 2pl, 3pl, 1- oder 2-/3-dimensional) berechnet und im zugehörigen Modellvergleich geprüft, welches Modell am besten zu den vorliegenden Daten passt. Das Modell mit der besten Passung wird für den jeweiligen MZP für die weiteren Untersuchungen verwendet.

6.2.1 Skalierbarkeit als Gesamtdatensatz

Es ist zu prüfen, ob die bisherige Betrachtung als Gesamtdatensatz zulässig und damit eine entsprechende gemeinsame Analyse und Skalierung möglich ist. Bereits in Kap. 5.5 wurden Unterschiede hinsichtlich der demographischen Variablen untersucht, wobei nur typische Kohortenunterschiede festgestellt werden konnten, die eine gemeinsame Betrachtung zulassen (siehe Kap. 5.5). Aber auch das Antwortverhalten der Probanden in den fachlichen Leistungstests muss untersucht werden. Dazu ist sowohl zu untersuchen, ob Haupterhebung und Nacherhebung der Studie FUNDAMENT signifikante Unterschiede aufweisen, als auch, ob sich die beiden Studien (FUNDAMENT und ALSTER) signifikant voneinander unterscheiden. Dazu wird das von Dammann (2016) vorgeschlagene Verfahren adaptiert: Zunächst erfolgt eine IRT-Skalierung (Rasch-Modell [1pl]) der Teilstichproben mit ›tam.mml‹ des R-Pakets ›TAM‹ (Robitzsch et al., 2022). Für den Vergleich der beiden Studien werden drei Modelle untersucht, die im Folgenden exemplarisch für MZP 2 erläutert werden: Modell 1 beinhaltet alle Ankeritems (zwischen FUNDAMENT und ALSTER) von MZP 2 mit den Probanden von FUNDAMENT, Modell 2 beinhaltet ebenfalls die Ankeritems, jedoch mit den Probanden von ALSTER und das dritte Modell beinhaltet die Ankeritems mit den Probanden aus beiden Studien. Die Varianzen werden durch die unterschiedlichen Stichproben beeinflusst, daher ist ein Modellvergleich mittels Likelihood-Ratio-Test nicht möglich (Dammann, 2016). Alternativ empfiehlt Dammann (2016) den Vergleich der Rangfolge der Itemschwierigkeiten, um eine Skalierbarkeit als Gesamtdatensatz zu ermöglichen. Die zuvor skalierten Itemschwierigkeiten der Modelle

werden mittels Produkt-Moment-Korrelation nach Pearson miteinander verglichen. Die Voraussetzungen (Intervallskalierung, linearer Zusammenhang zwischen den Variablen, bivariate Normalverteilung [Field et al., 2012] und keine Ausreißer in den Teilstichproben [Eid et al., 2017]) sind alle erfüllt, die Berechnung erfolgt mit der Funktion ›cor.test‹ aus dem ›R‹-Paket ›stats‹ (R Core Team, 2023). Die berechnete Effektstärke, der Pearson Korrelationskoeffizient (r), kann in drei Bereiche unterteilt werden:

- $|r| \geq 0.1$ - kleiner Effekt
- $|r| \geq 0.3$ - mittlerer Effekt
- $|r| \geq 0.5$ - großer Effekt (Cohen, 1992).

Eine hohe Effektstärke deutet auf gleiche Rangfolgen der Itemschwierigkeiten hin und lässt somit eine Betrachtung als Gesamtdatensatz zu (Dammann, 2016).

Aufgrund der geringen Probandenzahl (siehe Tab. 5.4) an MZP 1 wird der Empfehlung von Koller et al. (2015) gefolgt und auf die IRT-Skalierung an diesem MZP verzichtet.

Das beschriebene Verfahren wird auf den Teildatensatz der Studie FUNDAMENT angewendet, um signifikante Zusammenhänge zwischen der Haupt- und der Nacherhebung aufzuzeigen. Dabei werden die fachlichen Leistungstests der Technischen Mechanik und der Rechenfähigkeit getrennt betrachtet.

Die Effektstärken sind in Tab. 6.4 aufgelistet. Für die beiden Testinstrumente der Technischen Mechanik (FW und MF) ergeben sich für MZP 2 und 3 signifikante Ergebnisse mit hoher Effektstärke. Gleiches gilt für den vierten MZP, bei dem allerdings neben den Testinstrumenten FW und MF auch das Testinstrument TM_4 berücksichtigt wird. Da die hohen Effektstärken auf gleiche Rangfolgen der Itemschwierigkeiten hindeuten, ist eine Betrachtung als Gesamtdatensatz zulässig.

Auch das Testinstrument zur Rechenfähigkeit zeigt bei MZP 2 und 3 signifikante Zusammenhänge mit hohen Effektstärken. Anders verhält es sich jedoch an MZP 4, wo sich unter Berücksichtigung der Nacherhebung keine signifikanten Effekte zeigen. Somit können die Items

Tabelle 6.4
Vergleich der Pearson Korrelationskoeffizienten der Itemschwierigkeiten der Haupt- und Nacherhebung (FUNDAMENT)

	Modell 2	Modell 3
Technische Mechanik (FW, MF [und TM_4])		
MZP 2		
Modell 1	$.72^{***}$	$.96^{***}$
Modell 2		$.88^{***}$
MZP 3		
Modell 1	$.90^{***}$	$.98^{***}$
Modell 2		$.97^{***}$
MZP 4		
Modell 1	$.83^{***}$	$.99^{***}$
Modell 2		$.88^{***}$
Rechenfähigkeit (RF)		
MZP 2		
Modell 1	$.95^{***}$	$.99^{***}$
Modell 2		$.98^{***}$
MZP 3		
Modell 1	$.80^{***}$	$.95^{***}$
Modell 2		$.95^{***}$
MZP 4		
Modell 1	.31	$.99^{***}$
Modell 2		.44

Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4; MZP = Messzeitpunkt; RF = Rechenfähigkeit.
Modell 1 = Ankeritems - Probanden Haupterhebung;
Modell 2 = Ankeritems - Probanden Nacherhebung;
Modell 3 = Ankeritems - Probanden Haupterhebung und Nacherhebung.
*** $p < .001$.

der Rechenfähigkeit an MZP 4 nicht gemeinsam mit den Items der Haupterhebung betrachtet werden. Gründe hierfür könnten im Erhebungsformat (Online-Befragung) der Nacherhebung an MZP 4 liegen, auch ist die Probandenzahl (14 Probanden) deutlich geringer (siehe Tab. 5.4). Trotz der geringen Probandenzahl besteht diese Problematik beim fachlichen Leistungstest der Technischen Mechanik nicht. Da der Rechenfähigkeitstest den Abschluss der Online-Befragung bildet, könnte die Motivation der Probanden am Ende der Online-Befragung stark nachgelassen haben. Die 14 Teildatensätze der Rechenfähigkeit (FUNDAMENT MZP 4) werden von der Auswertung ausgeschlossen: Eine gemeinsame Skalierung ist wie erläutert nicht möglich, eine getrennte Skalierung aufgrund der geringen Probandenzahl laut Koller et al. (2015) nicht empfehlenswert (siehe Kap. 4.1).

Das gleiche Verfahren wird ebenfalls für den Vergleich der Datensätze der Studien FUNDAMENT und ALSTER angewendet. Auch hier werden die fachlichen Leistungstests (Technische Mechanik und Rechenfähigkeit) separat voneinander betrachtet. Die Auflistung der Effektstärken findet sich in Tab. 6.5. Die Ergebnisse des Testinstruments der Rechenfähigkeit werden für MZP 4 aufgrund der Kodierungsprobleme in ALSTER nicht berücksichtigt (siehe Kap. 5.1).

Der fachliche Leistungstest der Technischen Mechanik zeigt an allen drei MZP hohe Korrelationen zwischen den beiden Studien. Eine gemeinsame Betrachtung und Skalierung als Gesamtdatensatz ist daher zulässig.

Zwischen den Modellen 1 und 2 ist jedoch kein signifikanter Zusammenhang in Bezug auf die Rechenfähigkeit (MZP 2 und 3) erkennbar, d.h. die Itemschwierigkeiten besitzen nicht die gleiche Rangfolge und eine gemeinsame Betrachtung ist somit nicht zulässig. Da die fachlichen Leistungstests der Technischen Mechanik diese Problematik nicht zeigen, lässt sich schlussfolgern, dass die unterschiedlichen Antwortformate der Items der Rechenfähigkeit (siehe Kap. 5.2) der Auslöser sind. Die Daten zur Rechenfähigkeit der MZP 2 und 3 aus ALSTER verbleiben im Datensatz, müssen aber wie beschrieben separat skaliert werden, die Probandenzahlen (MZP 2=193 bzw. MZP 3=184) sind dafür ausreichend groß.

Tabelle 6.5
Vergleich der Pearson Korrelationskoeffizienten der Itemschwierigkeiten FUNDAMENT und ALSTER

	Modell 2	Modell 3
Technische Mechanik (FW und MF)		
MZP 2		
Modell 1	$.83^{***}$	$.97^{***}$
Modell 2		$.94^{***}$
MZP 3		
Modell 1	$.79^{***}$	$.94^{***}$
Modell 2		$.95^{***}$
MZP 4		
Modell 1	$.70^{***}$	$.96^{***}$
Modell 2		$.87^{***}$
Rechenfähigkeit (RF)		
MZP 2		
Modell 1	.39	$.81^{**}$
Modell 2		$.85^{***}$
MZP 3		
Modell 1	.32	.61
Modell 2		$.94^{**}$

Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; MZP = Messzeitpunkt; RF = Rechenfähigkeit.
Modell 1 = Ankeritems - Probanden FUNDAMENT;
Modell 2 = Ankeritems - Probanden ALSTER;
Modell 3 = Ankeritems - Probanden FUNDAMENT und ALSTER.
** $p < .01$; *** $p < .001$.

6.2.2 Modellpassung

Im vorangegangenen Kapitel (Kap. 6.2.1) wurde die Skalierbarkeit als Gesamtdatensatz untersucht. Innerhalb der Studie FUNDAMENT ist eine gemeinsame Betrachtung von Haupt- und Nacherhebung zulässig. Eine gemeinsame Skalierung der beiden Studien (FUNDAMENT und ALSTER) ist ebenfalls möglich, allerdings mit der Einschränkung, dass die Testinstrumente zur Erfassung der Rechenfähigkeit nicht die gleiche Rangfolge bei der Prüfung der Itemschwierigkeiten aufweisen und daher nicht gemeinsam skaliert werden dürfen. Deswegen werden bei der IRT-Skalierung die fachlichen Leistungstests der Technischen Mechanik und der Rechenfähigkeit getrennt skaliert. Während für die Testinstrumente der Technischen Mechanik eine gemeinsame Skalierung mit Verankerung der MZP beider Studien erfolgt, werden die Testinstrumente der Rechenfähigkeit separat skaliert und über die MZP verankert (FUNDAMENT: MZP 2 bis 4; ALSTER: MZP 2 und 3).

In einem ersten Schritt werden die MZP (ohne Verankerung) Rasch skaliert und die Modellpassung anhand der Items überprüft. Items, die einen problematischen wMNSQ-Infit und/oder wMNSQ-Outfit aufweisen, werden entsprechend aus der Analyse ausgeschlossen. Anschließend werden die finalen Datensätze wie zuvor beschrieben über die MZP verankert, um eine Beschreibung und Darstellung von Entwicklungen anhand einer gemeinsamen Skala zu ermöglichen. Ein Modellvergleich soll dann klären, welches IRT-Modell (1pl, 2pl, 3pl, 1- oder 2-/3-dimensional) am besten zu den Daten passt.

In diesem Kapitel werden die Fit-Werte (wMNSQ-Infit und wMNSQ-Outfit) der skalierten Items überprüft. Zunächst werden die Testinstrumente der Technischen Mechanik geprüft, wobei die Daten des Gesamtdatensatzes verwendet und die MZP 2-4 separat betrachtet werden. Anschließend erfolgt die Überprüfung der Testinstrumente der Rechenfähigkeit, die jedoch sowohl für FUNDAMENT und ALSTER als auch die jeweiligen MZP getrennt analysiert werden.

Bei der Prüfung werden die in Kap. 4.1.2 genannten Grenzen für den wMNSQ von 0.8 bis 1.2 angewendet. Für den Fall, dass ein Item Werte für den wMNSQ-Infit und/oder wMNSQ-Outfit außerhalb dieser Grenzen aufweist, wird als weiterer Fit-Wert das gewichtete t berücksichtigt.

Die Grenzen liegen hier zwischen -2 und 2. Hat ein Item zusätzlich einen gewichteten t-Wert außerhalb dieser Grenzen, wird das Item ausgeschlossen. Ein Ausschluss erfolgt nur, wenn beide Kriterien erfüllt sind. Ein gewichteter t-Wert außerhalb dieser Grenzen allein ist nicht ausreichend, da nach Wilson (2023) dieser Wert bei großen Stichproben bereits für viele, wenn nicht sogar für alle Items signifikant wird. Anschließend erfolgt eine erneute Skalierung ohne dieses auffällige Item. Dieser Prozess wird so lange durchgeführt, bis keine auffälligen Items mehr vorhanden sind. Die Skalierung erfolgt in ›R‹ (R Core Team, 2023) mit dem Paket ›TAM‹ (Robitzsch et al., 2022) und der Funktion ›tam.fit‹.

Technische Mechanik Für die Testinstrumente der Technischen Mechanik an MZP 2 liegen die Werte des wMNSQ-Infit in einem akzeptablen Bereich $[0.867; 1.153]$, während vereinzelte Werte des wMNSQ-Outfit $[0.705; 1.414]$ außerhalb der Grenzen liegen. Da auch die gewichteten t-Werte jeweils außerhalb der akzeptablen Grenzen liegen, mussten schrittweise vier Items ausgeschlossen werden, sodass an MZP 2 noch 83 Items (wMNSQ-Infit $[0.862; 1.177]$ und wMNSQ-Outfit $[0.693; 1.198]$) im Datensatz verbleiben. Ein wMNSQ-Outfit liegt mit 0.693 außerhalb der Grenzen, dieses Item verbleibt jedoch im Datensatz, da das gewichtete t mit -1.645 innerhalb der Grenzen liegt.

An MZP 3 wurden insgesamt 33 Items ausgeschlossen. Diese Zahl erscheint auf den ersten Blick recht hoch, allerdings wurde auch eine sehr hohe Anzahl von Items (118) eingesetzt, sodass diese hohe Zahl durchaus zu erwarten war. Vor allem vor dem Hintergrund, dass in dieser Arbeit die eher als restriktiv zu bezeichnenden Grenzen von PISA verwendet werden. Nach Ausschluss der Items, die vor allem aufgrund auffälliger wMNSQ-Outfit-Werte (und der entsprechenden gewichteten t-Werte) ausgeschlossen wurden, liegen die Werte des wMNSQ-Infit $[0.865; 1.147]$ innerhalb der vorgegebenen Grenzen. Die Werte des wMNSQ-Outfit liegen im Bereich $[0.810; 1.261]$, hier gibt es ein Item, das einen Wert außerhalb der Grenzen aufweist, jedoch liegt das gewichtete t mit 1.798 innerhalb der Grenzen und somit verbleibt das Item im Datensatz. An MZP 3 verbleiben 85 Items im Datensatz.

Eine ähnliche Ausgangslage ergibt sich an MZP 4. Vorrangig auf-

grund auffälliger wMNSQ-Outfit-Werte (und der entsprechenden gewichteten t-Werte) wurden 56 Items ausgeschlossen. Dieser Wert ist noch höher als die 32 ausgeschlossenen Items an MZP 3, aber auch die ursprüngliche Anzahl der Items ist mit 135 Items höher. Nach Ausschluss der 55 Items bleiben 79 Items im Datensatz. Wiederum gibt es zwei Items, die einen wMNSQ-Outfit-Wert außerhalb der Grenzen haben (1.207 und 1.259), diese beiden Items verbleiben aber im Datensatz, da die gewichteten t-Werte mit 1.612 und 1.480 innerhalb der Grenzen liegen. Insgesamt liegt an MZP 4 der wMNSQ-Infit im Bereich von $[0.828; 1.189]$ und der wMNSQ-Outfit im Bereich $[0.801; 1.259]$.

Rechenfähigkeit Wie bereits erläutert, wird die Rechenfähigkeit für die Studien FUNDAMENT und ALSTER getrennt berechnet. Während für FUNDAMENT die MZP 2 bis 4 betrachtet werden, stehen für ALSTER nur die MZP 2 und 3 zur Verfügung. Zunächst werden die Fit-Werte der Testinstrumente der Rechenfähigkeit von FUNDAMENT untersucht. An MZP 2 weisen fünf Items auffällige wMNSQ-Outfit-Werte auf, bei denen auch die gewichteten t-Werte außerhalb der Grenzen liegen. Nach Ausschluss dieser fünf Items bleiben im Datensatz 20 Items mit einem wMNSQ-Infit im Bereich $[0.862; 1.090]$ und einem wMNSQ-Outfit im Bereich $[0.818; 1.187]$.

Zwei Items an MZP 3 weisen sowohl für den wMNSQ-Infit als auch für den wMNSQ-Outfit auffällige Werte auf, jeweils mit gewichteten t-Werten außerhalb der Grenzen. Diese beiden Items wurden aus dem Datensatz entfernt, sodass 12 Items mit einem wMNSQ-Infit im Bereich $[0.861; 1.138]$ und wMNSQ-Outfit im Bereich $[0.802; 1.176]$ verbleiben.

Der wMNSQ-Outfit zeigt an MZP 4 für zwei Items auffällige Werte, die bei entsprechend gewichteten t-Werten außerhalb der Grenzen zum Ausschluss der Items aus dem Datensatz führen. Insgesamt bleiben somit an diesem MZP 12 Items im Datensatz (wMNSQ-Infit $[0.934; 1.061]$ und wMNSQ-Outfit $[0.866; 1.215]$). Bei einem dieser Items ist der Wert für den wMNSQ-Outfit noch zu hoch (1.215), aber da der gewichtete t-Wert innerhalb der Grenzen liegt, bleibt das Item im Datensatz.

Die Fit-Werte für die Testinstrumente der Rechenfähigkeit von ALSTER sind an MZP 2 für 13 Items auffällig (hauptsächlich wMNSQ-Outfit, aber auch auch zwei wMNSQ-Infit sowie gewichtete t-Werte

außerhalb der Grenzen). Diese Items werden aus dem Datensatz ausgeschlossen, sodass 14 Items im Datensatz verbleiben (wMNSQ-Infit $[0.901; 1.131]$ und wMNSQ-Outfit $[0.820; 1.199]$).

Am dritten MZP zeigen insgesamt 20 Items Auffälligkeiten (wMNSQ-Outfit und gewichtete t-Werte außerhalb der Grenzen) und werden entsprechend aus dem Datensatz ausgeschlossen. Es verbleiben 10 Items im Datensatz mit einem wMNSQ-Infit im Bereich von $[0.896; 1.115]$ und einem wMNSQ-Outfit im Bereich von $[0.809; 1.143]$.

Die nach der Modellpassung im Datensatz verbleibenden Items und Ankeritems sind in Tab. 6.6 aufgelistet.

Tabelle 6.6
Itemanzahl nach Modellpassung

	MZP 2	MZP 3	MZP 4
Technische Mechanik	83	85	79
Anker		54	57
Rechenfähigkeit			
FUNDAMENT	20	12	12
Anker		10	10
ALSTER	14	10	
Anker		6	

Anmerkungen. MZP = Messzeitpunkt.
Die Anzahl der Ankeritems bezieht sich auf die Verankerung zum vorherigen Messzeitpunkt.

Nach der erfolgten Modellpassung fällt auf, dass der wMNSQ-Outfit häufig außerhalb der akzeptablen Grenzen liegt. Diese Items beschreiben Personen unzureichend, deren Personenfähigkeit weit von der Itemschwierigkeit entfernt ist. Sofern diese Personen für die Messung weniger relevant sind, könnten die Items im Datensatz verbleiben. Da jedoch alle Probanden im Datensatz für diese Untersuchung wichtig sind, werden die entsprechenden Items ausgeschlossen. Wie bereits erwähnt, wurden in dieser Untersuchung die restriktiven Grenzen von PISA verwendet, was zwangsläufig zu einer höheren Anzahl ausgeschlossener Items führt. Da jedoch, wie in Tab. 6.6 ersichtlich, die Anzahl der Items nach Ausschluss der auffälligen Items immer noch groß ist, ist dieses Vorgehen vertretbar.

6.2.3 Modellvergleiche (IRT)

Die nach der Modellpassung im Datensatz verbleibenden Items bilden den finalen Datensatz, der in diesem Kapitel skaliert wird. Wie bereits in Kap. 6.2.1 beschrieben, werden zunächst die Testinstrumente der Technischen Mechanik für beide Studien gemeinsam skaliert. Anschließend wird die Skalierung der Testinstrumente der Rechenfähigkeit durchgeführt, jedoch für beide Studien getrennt. Die Skalierung erfolgt über die Verankerung der Ankeritems, die bereits am jeweiligen vorangegangenen MZP verwendet wurden. Welches IRT-Modell (1pl, 2pl, 3pl, 1- oder 2-/3-dimensional) am besten zu den vorliegenden Daten passt, wird durch Modellvergleiche untersucht, wobei die drei Informationskriterien Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) und Consistent Akaike Information Criterion (CAIC), die EAP-Reliabilität und der χ^2-*Anpassungstest* ($\alpha = .05$) herangezogen werden.

6.2.3.1 Technische Mechanik

Zunächst werden die drei MZP der Testinstrumente der Technischen Mechanik für beide Studien gemeinsam skaliert. Die Berechnung erfolgt in »R« (R Core Team, 2023) mit dem Paket ›TAM‹ (Robitzsch et al., 2022) und den Funktionen ›tam.mml‹, ›tam.mml.2pl‹ bzw. ›tam.mml.3pl‹. Für die Schätzung der Personenfähigkeiten wird die Funktion ›tam.wle‹ verwendet, die den *Weighted Maximum Likelihood Estimates*-Schätzer verwendet.

Für jeden MZP werden die folgenden vier verschiedenen Modelle skaliert:

- Modell 1 = Rasch-Modell (1pl, 1-Dim)
- Modell 2 = Rasch-Modell (1pl, 2/3-Dim)
- Modell 3 = Birnbaum-Modell (2pl, 1-Dim)
- Modell 4 = Birnbaum-Modell (3pl, 1-Dim).

Während für die Modelle 1 bis 3 alle Parameter geschätzt werden, wird für Modell 4 der Rateparameter für jedes Item auf $\gamma_j = .25$ gesetzt. Die entsprechenden Modelle werden für alle MZP berechnet, wobei für

die Ankeritems (MZP 3 und 4) die Itemschwierigkeiten des vorherigen MZP fixiert werden.

Die Überprüfung der Items auf DIF erfolgt anhand von vier Verfahren zur Ermittlung von Item-DIF. Sofern alle vier Verfahren bei einem Item DIF erkennen, wird das Item ausgeschlossen. Bei den Verfahren handelt es sich um ein uniformes Verfahren mit *Mantel-Haenszel* (Holland & Thayer, 1988), ein non-uniformes Verfahren mit der *Logistischen Regression* (Swaminathan & Rogers, 1990) und zwei IRT-Verfahren mit dem *Lord's* χ^2*-Test* (Lord, 1980) und *Raju's Flächenmethode* (Raju, 1990). Es werden jeweils zwei Gruppenmerkmale auf DIF untersucht, wobei Geschlecht und Muttersprache (Deutsch) als sinnvolle Kriterien erachtet werden (Dammann, 2016). Keines der in den Testinstrumenten der Technischen Mechanik verwendeten Items weist in allen vier Verfahren DIF auf, sodass keines der Items ausgeschlossen werden muss.

MZP 2 Für den zweiten MZP stehen 83 Items zur Verfügung (siehe Tab. 6.6).

In Tab. 6.7 sind die Kennwerte der 1pl-Rasch-Modelle 1- (Modell 1) und 2-dimensional (Modell 2) aufgelistet. Beim 2-dimensionalen Modell (Modell 2) ist die ersten Dimension das Fachwissen und die zweite Dimension die Modellierungsfähigkeit. Zunächst wird die Anzahl der in der Skalierung berücksichtigten Probanden ($n_{\text{Probanden}}$) angegeben, in allen Modellen am zweiten MZP sind dies 484 Probanden. Die Genauigkeit der Anpassung an das jeweilige IRT-Modell kann durch einen statistischen Wert der logarithmischen Likelihood-Funktion (LogLike.) und der Deviance ausgedrückt werden (siehe Kap. 4.1.4). Beide Werte werden zur Berechnung des *Likelihood-Ratio-Test* und der Informationskriterien benötigt.

Die drei Informationskriterien AIC, BIC und CAIC sind im Modell 1 alle etwas niedriger als im 2-dimensionalen Modell, sodass nach diesen Kriterien das erste Modell für die vorliegenden Daten besser geeignet wäre. Eine höhere Anzahl geschätzter Parameter deutet auf ein komplexeres Modell hin. Da im 2-dimensionalen Modell für jede Dimension Parameter geschätzt werden müssen, ist dieser Wert entsprechend höher. Für Dimension 1 (in Modell 1 gibt es nur eine Dimension) bzw. Dimension 2 wird die Anzahl der berücksichtigten Items angege-

Tabelle 6.7
Kennwerte IRT-Modelle (1pl, 1- und 2-dimensional) Testinstrument der Technischen Mechanik an MZP 2

	Modell 1	Modell 2
$n_{\text{Probanden}}$	484	484
LogLike.	−17 078.83	−17 118.27
Deviance	34 157.67	34 236.53
AIC	34 324	34 409
BIC	34 671	34 768
CAIC	34 754	34 854
Geschätzte Parameter	84	86
Dimension 1		
n_{Items}	83	37
EAP-Reliabilität	0.753	0.718
wMNSQ-Infit	[0.862; 1.177]	[0.918; 1.175]
wMNSQ-Outfit	[0.843; 1.198]	[0.801; 1.199]
Varianz	0.304	0.326
Dimension 2		
n_{Items}		46
EAP-Reliabilität		0.707
wMNSQ-Infit		[0.861; 1.131]
wMNSQ-Outfit		[0.834; 1.185]
Varianz		0.280

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 1 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Rasch-Modell (1pl, 2-Dim): Dimension 1 = Fachwissen, Dimension 2 = Modellierungsfähigkeit.

ben. Während in Modell 1 alle Items auf der ersten Dimension liegen, verteilen sich diese 82 Items im zweiten Modell auf Dimension 1 (Fachwissen - 37 Items) und Dimension 2 (Modellierungsfähigkeit - 46 Items). Die EAP-Reliabilität (Expected a Posteriori-Reliabilität) ist nach Wess et al. (2021) vergleichbar mit Cronbach´s α und kann entsprechend interpretiert werden (siehe Kap. 4.1.2). Alle erzielten Werte können als akzeptabel angesehen werden. Modell 1 erreicht mit .751 die höchste Reliabilität, was allerdings auch auf die höhere Anzahl der Items im Vergleich zu den einzelnen Dimensionen des zweiten Modells zurückzuführen ist. Als zusätzliche Information sind die kleinsten und größten wMNSQ-Infit/Outfit-Werte sowie die Varianz angegeben. Auffällige wMNSQ-Infit/Outfit-Werte sind entsprechend gekennzeichnet. Sofern nicht anders beschrieben, verbleiben diese Items im Datensatz, da die gewichteten t-Werte innerhalb der Grenzen liegen.

Zusätzlich zu den Informationskriterien wird die Funktion ›anova‹ aus dem R-Paket ›lavaan‹ (Rosseel, 2012) verwendet, um einen Modellvergleich durchzuführen. Diese Funktion führt den χ^2-*Anpassungstest* (siehe Kap. 4.1.4) durch. Bei einem signifikanten Ergebnis passt das Modell mit mehr freigeschätzten Parametern besser zu den Daten als das Modell mit weniger frei geschätzten Parametern. Bei einem nicht signifikanten Ergebnis ist aus Gründen der Sparsamkeit das Modell mit der geringeren Anzahl an frei geschätzten Parametern zu bevorzugen. Der Modellvergleich zwischen Modell 1 und Modell 2 zeigt kein signifikantes Ergebnis ($\chi^2(3) = -78.87$, $p > .999$) und bestätigt damit die Aussage der Informationskriterien. Das 1-dimensionale Modell passt besser zu den Daten und ist daher dem 2-dimensionalen Modell vorzuziehen.

Neben der Skalierung als 1pl-Rasch-Modell wird eine Berechnung des 2pl- (Modell 3) und 3pl-Modells (Modell 4) nach Birnbaum durchgeführt. In beiden Modellen werden die Trennschärfeparameter frei geschätzt. Für Modell 4 (Birnbaum-Modell, 3pl, 1-Dim) wird der Rateparameter für jedes Item auf $\gamma_j = .25$ gesetzt. Dieser Wert ergibt sich aus der Tatsache, dass es sich um Multiple-Choice-Items mit vier Antwortmöglichkeiten handelt, sodass die Wahrscheinlichkeit, die richtige Antwort zu erraten, 25% beträgt.

Die Kennwerte der beiden Modelle sind in Tab. 6.8 dargestellt. Mo-

dell 3 zeigt für alle Informationskriterien geringfügig kleinere Werte und damit eine bessere Passung der Daten. Auch die EAP-Reliabilität kann bei Modell 3 als gut bezeichnet werden, während der Wert bei Modell 4 nur als akzeptabel eingestuft werden kann. Diese Aussagen lassen sich auch mit dem χ^2-*Anpassungstest* bestätigen, der kein signifikantes Ergebnis liefert ($\chi^2(1) = -15$, $p > .999$). Daher ist das Modell mit der geringeren Anzahl frei zu schätzender Parameter (Modell 3 - 2pl-Birnbaum-Modell) zu bevorzugen.

Tabelle 6.8
Kennwerte IRT-Modelle (2pl und 3pl) - Testinstrument der Technischen Mechanik an MZP 2

	Modell 3	Modell 4
$n_{\text{Probanden}}$	484	484
LogLike.	−16 887.52	−16 895.02
Deviance	33 775.04	33 790.04
AIC	34 107	34 124
BIC	34 801	34 822
CAIC	34 967	34 989
Geschätzte Parameter	166	167
n_{Items}	83	83
EAP-Reliabilität	0.813	0.789
wMNSQ-Infit	[0.978 ; 1.021]	[0.718*; 1.105]
wMNSQ-Outfit	[0.727*; 1.229*]	[0.713*; 1.251*]
Varianz	1.000	0.424

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 3 = Birnbaum-Modell (2pl, 1-Dim);
Modell 4 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

Außerdem ist zu prüfen, welches Modell (1 oder 3) vorzuziehen ist. Zur besseren Veranschaulichung sind die entsprechenden Daten in Tab. 6.9 aufgelistet. Ein Vergleich der EAP-Reliabilitäten zeigt einen

höheren und damit besseren Wert für Modell 3. Die Informationskriterien hingegen sind nicht eindeutig. Während Modell 3 einen niedrigeren AIC aufweist, sind die beiden anderen Informationskriterien bei Modell 1 niedriger. Der χ^2-*Anpassungstest* zeigt ein signifikantes Ergebnis ($\chi^2(82) = 481.53$, $p < .001$), welches dementsprechend das dritte Modell mit der höheren Anzahl geschätzter Parameter bevorzugt. Da nur BIC und CAIC Modell 1 bevorzugen, wird Modell 3 als das für die vorliegenden Daten am besten geeignete Modell angesehen und im Folgenden weiter verwendet.

Tabelle 6.9
Kennwerte IRT-Modelle (1pl und 2pl) - Testinstrument der Technischen Mechanik an MZP 2

	Modell 1	Modell 3
AIC	34 324	34 107
BIC	34 671	34 801
CAIC	34 754	34 967
Geschätzte Parameter	84	166
EAP-Reliabilität	0.753	0.813

Anmerkungen. MZP = Messzeitpunkt; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; EAP-Reliabilität = Expected a Posteriori-Reliabilität.
Modell 1 = Rasch-Modell (1pl, 1-Dim);
Modell 3 = Birnbaum-Modell (2pl, 1-Dim).

Tab. 6.10 listet die zentralen Kennwerte der berechneten Itemschwierigkeiten (β) und Personenfähigkeiten (θ) des 2pl-Birnbaum-Modells auf. Neben der Stichprobengröße (n) wird der jeweilige Mittelwert (M), das Minimum (Min) und das Maximum (Max) angegeben. An MZP 2 beträgt die mittlere Itemschwierigkeit $\beta_M = 0.21$, die Personenfähigkeit $\theta_M = -0.01$.

Eine Wright Map visualisiert diese Kennwerte. Die Wright Map der Skalierung des zweiten MZP der Testinstrumente der Technischen Mechanik ist in Abb. 6.6 dargestellt.

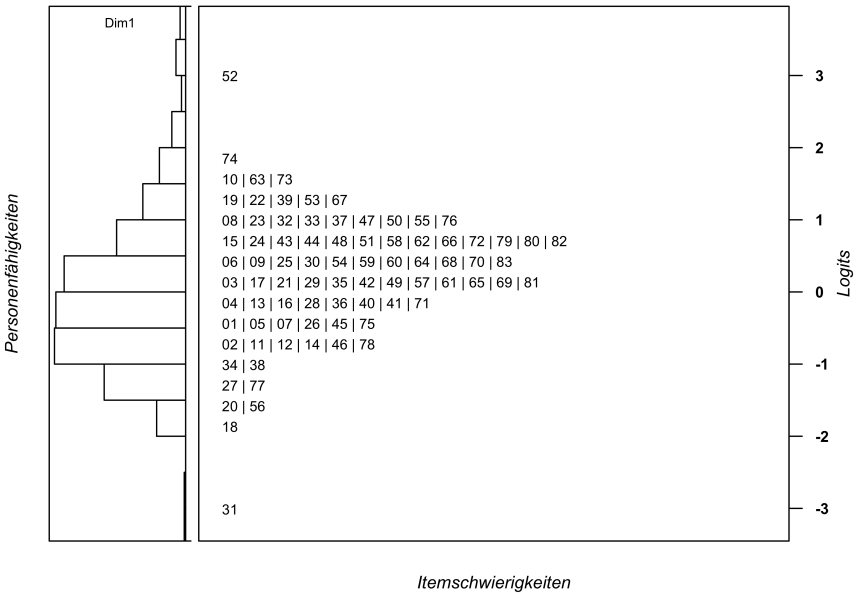
Auf der linken Seite ist ein Histogramm der Personenfähigkeiten der 484 Probanden im Datensatz dargestellt, auf der rechten Seite die

Tabelle 6.10
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 2pl) - Testinstrument der Technischen Mechanik an MZP 2

	MZP 2
Itemschwierigkeiten	
n_{Items}	83
β_M	0.21
β_{Min}	−2.93
β_{Max}	2.95
Personenfähigkeiten	
$n_{\text{Probanden}}$	484
θ_M	−0.01
θ_{Min}	−3.06
θ_{Max}	4.18

Anmerkungen. MZP = Messzeitpunkt; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Abbildung 6.6
Wright Map - Testinstrument der Technischen Mechanik an MZP 2



Itemschwierigkeiten, wobei die Zahlen jeweils für ein Item stehen (die Nummerierung entspricht der Reihenfolge im Datensatz). Auf der Y-Achse ist die Logit-Skala für den Wertebereich von -3 bis 3 aufgetragen.

MZP 3 Die Items von MZP 2, die auch am dritten MZP verwendet werden, sind Ankeritems. Für diese Ankeritems wird die geschätzte Itemschwierigkeit des 2pl-Birnbaum-Modells (MZP 2) übernommen und als Itemschwierigkeit für die entsprechenden Items des dritten MZP festgelegt. Dies sind 54 Items (siehe Tab. 6.6), sodass weniger Parameter für den dritten MZP geschätzt werden müssen. Für den MZP 3 wird die gleiche Vorgehensweise wie für den zweiten MZP angewendet, es werden also wieder vier Modelle mit den Daten berechnet und miteinander verglichen (siehe Kap. 6.2.3.1).

Die Kennwerte der 1pl-Rasch-Modelle 1- (Modell 1) und 2-dimensional (Modell 2) sind in Tab. 6.11 aufgelistet. Analog zum vorherigen MZP ist beim 2-dimensionalen Modell (Modell 2) die erste Dimension das Fachwissen und die zweite Dimension die Modellierungsfähigkeit.

299 Probanden werden bei der Skalierung der Modelle berücksichtigt. Ein Vergleich der drei Informationskriterien zeigt, dass Modell 1 minimal niedrigere Werte erzielt als das zweite Modell und somit eine bessere Passung der Daten gewährleistet. Auch die EAP-Reliabilitäten unterstützen diese Einschätzung, da das erste Modell hier einen höheren Wert erzielt. Insgesamt liegen die EAP-Reliabilitäten beider Modelle in einem guten Bereich. Auch der χ^2-*Anpassungstest* bevorzugt Modell 1 ($\chi^2(3) = 3.870$, $p = .276$).

Bei der Berechnung des 2pl- (Modell 3) und 3pl-Modells (Modell 4) nach Birnbaum werden die Trennschärfeparameter frei geschätzt, der Rateparameter (Modell 4) wird für jedes Item auf $\gamma_j = .25$ gesetzt. In Tab. 6.12 sind die Kennwerte der beiden Modelle aufgeführt.

Die Werte des AIC, BIC und CAIC sind bei Modell 3 etwas niedriger, sodass Modell 3 aufgrund der Informationskriterien dem Modell 4 vorzuziehen ist. Dies zeigt sich auch bei der EAP-Realiabilität, bei der Modell 3 einen etwas höheren Wert aufweist. Auch das dritte Kriterium des Modellvergleichs, der χ^2-*Anpassungstest*, bevorzugt Modell 3 ($\chi^2(1) = 269.41$, $p > .999$).

Wie bei MZP 2 ist Modell 3 den anderen drei Modellen überle-

Tabelle 6.11
Kennwerte IRT-Modelle (1pl, 1- und 2-dimensional) Testinstrument der Technischen Mechanik an MZP 3

	Modell 1	Modell 2
$n_{\text{Probanden}}$	299	299
LogLike.	−11 793.09	−11 791.16
Deviance	23 586.18	23 582.31
AIC	23 654	23 656
BIC	23 780	23 793
CAIC	23 814	23 830
Geschätzte Parameter	34	37
Dimension 1		
n_{Items}	85	36
EAP-Reliabilität	0.865	0.830
wMNSQ-Infit	[0.778*; 1.197]	[0.781*; 1.366*]
wMNSQ-Outfit	[0.816 ; 1.266*]	[0.817 ; 1.319*]
Varianz	0.615	0.634
Dimension 2		
n_{Items}		49
EAP-Reliabilität		0.849
wMNSQ-Infit		[0.802 ; 1.180]
wMNSQ-Outfit		[0.727*; 1.248*]
Varianz		0.594

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 1 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Rasch-Modell (1pl, 2-Dim): Dimension 1 = Fachwissen, Dimension 2 = Modellierungsfähigkeit.
* = Ein Item liegt außerhalb der Grenzen.

Tabelle 6.12
Kennwerte IRT-Modelle (2pl und 3pl) - Testinstrument der Technischen Mechanik an MZP 3

	Modell 3	Modell 4
$n_{\text{Probanden}}$	299	299
LogLike.	−11 628.70	−11 763.40
Deviance	23 257.40	23 526.81
AIC	23 493	23 765
BIC	23 930	24 205
CAIC	24 048	24 324
Geschätzte Parameter	118	119
n_{Items}	85	85
EAP-Reliabilität	0.872	0.861
wMNSQ-Infit	[0.891 ; 1.241*]	[0.741*; 1.229*]
wMNSQ-Outfit	[0.863 ; 1.374*]	[0.715*; 1.202*]
Varianz	1.000	0.716

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 3 = Birnbaum-Modell (2pl, 1-Dim);
Modell 4 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt
* = Ein Item liegt außerhalb der Grenzen.

gen (Tab. 6.13). Im Vergleich zu Modell 1 weist Modell 3 einen etwas geringeren Wert für das AIC auf, während die beiden anderen Informationskriterien bei Modell 1 etwas kleiner sind. Die EAP-Reliabilität ist etwas höher als bei Modell 1 und auch der χ^2-*Anpassungstest* zeigt ein signifikantes Ergebnis ($\chi^2(84) = 382.78$, $p < .001$) zugunsten von Modell 3. Somit ist das dritte Modell (2pl-Birnbaum-Modell) für die vorliegenden Daten am besten geeignet und wird im Folgenden weiter verwendet.

Tabelle 6.13
Kennwerte IRT-Modelle (1pl und 2pl) - Testinstrument der Technischen Mechanik an MZP 3

	Modell 1	Modell 3
AIC	23 654	23 493
BIC	23 780	23 930
CAIC	23 814	24 048
Geschätzte Parameter	34	118
EAP-Reliabilität	0.865	0.872

Anmerkungen. MZP = Messzeitpunkt; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; EAP-Reliabilität = Expected a Posteriori-Reliabilität.
Modell 1 = Rasch-Modell (1pl, 1-Dim);
Modell 3 = Birnbaum-Modell (2pl, 1-Dim).

Die berechnete mittlere Itemschwierigkeit für MZP 3 (siehe Tab. 6.14) beträgt $\beta_M = 0.37$ und ist damit etwas höher als für MZP 2 ($\beta_M = 0.21$). Die mittlere Personenfähigkeit liegt mit $\theta_M = 0.89$ deutlich über dem Wert von MZP 2 ($\theta_M = -0.01$). Somit kann für die Technische Mechanik von einem Kompetenzerwerb im ersten Fachsemester gesprochen werden.

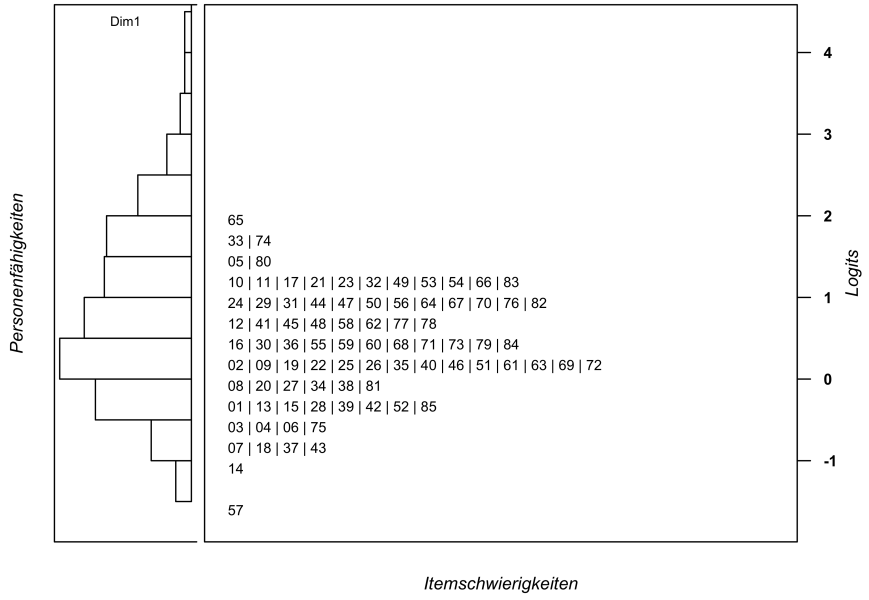
In Abb. 6.7 ist die Wright Map der Testinstrumente der Technischen Mechanik an MZP 3 visualisiert. Zu beachten ist, dass die Logit-Skala einen anderen Wertebereich zeigt als die Wright Map für MZP 2 (Abb. 6.6).

Tabelle 6.14
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 2pl) - Testinstrument der Technischen Mechanik an MZP 3

	MZP 3
Itemschwierigkeiten	
n_{Items}	85
β_M	0.37
β_{Min}	−1.5
β_{Max}	1.87
Personenfähigkeiten	
$n_{\text{Probanden}}$	299
θ_M	0.89
θ_{Min}	−1.41
θ_{Max}	6.05

Anmerkungen. MZP = Messzeitpunkt; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Abbildung 6.7
Wright Map - Testinstrument der Technischen Mechanik an MZP 3



MZP 4 Die Itemschwierigkeiten des 2pl-Birnbaum-Modells aus MZP 3 werden für die Ankeritems übernommen, sodass für die entsprechenden Items die Itemschwierigkeiten an MZP 4 fixiert sind. Dies ist bei insgesamt 57 Items der Fall (siehe Tab. 6.6). Wie bei den beiden vorherigen MZP werden zunächst vier Modelle gebildet. Lediglich Modell 2 unterscheidet sich, da hier kein 2-dimensionales, sondern ein 3-dimensionales Modell berechnet wird.

Tab. 6.15 zeigt die Kennwerte der 1pl-Rasch-Modelle 1- (Modell 1) und 3-dimensional (Modell 2). Die Dimensionen des 3-dimensionalen Modells (Modell 2) sind in der ersten Dimension das Fachwissen, in der zweiten die Modellierungsfähigkeit und in der dritten die Items der Technischen Mechanik 2 (TM_4). Da bei den Items der TM_4 keine klare Trennung zwischen Fachwissen und Modellierungsfähigkeit besteht, werden diese Items einer eigenen Dimension zugeordnet.

Im Gegensatz zu den vorherigen MZP zeigt das mehrdimensionale Modell (Modell 2) eine bessere Passung zu den vorliegenden Daten. Alle drei Informationskriterien sind im 3-dimensionalen Modell (Modell 2) niedriger als im 1-dimensionalen Modell. Die EAP-Reliabilitäten liegen dagegen mit .791 bis .847 etwas niedriger als bei dem 1-dimensionalen Modell (.864). Während die EAP-Reliabilitäten für Modell 1 und die Dimensionen 1 (Fachwissen) und 2 (Modellierungsfähigkeit) als gut bezeichnet werden können, liegt der Wert für die dritte Dimension (Items der Technischen Mechanik 2 [TM4]) in einem akzeptablen Bereich. Dieser Unterschied kann durch die geringe Anzahl von nur noch 9 Items im Datensatz erklärt werden. Die entsprechenden Items (ursprünglich 21) wurden im Rahmen von FUNDAMENT neu entwickelt und konnten aufgrund mangelnder Probandenzahl nicht ausreichend pilotiert werden.

Der χ^2-*Anpassungstest* zeigt ein signifikantes Ergebnis ($\chi^2(7) = 15.771$, $p = .027$), sodass sowohl auf dieser Basis als auch nach den Informationskriterien Modell 2 dem ersten Modell vorzuziehen ist.

Die Kennwerte des 2pl- (Modell 3) und 3pl-Modells (Modell 4) nach Birnbaum (1-dimensional) sind in Tab. 6.12 aufgeführt. Analog zu den vorhergehenden MZP werden die Trennschärfeparameter frei geschätzt und für das 3pl-Modell (Modell 4) wird der Rateparameter für jedes Item auf $\gamma_j = .25$ gesetzt.

Tabelle 6.15
Kennwerte IRT-Modelle (1pl, 1- und 3-dimensional) Testinstrument der Technischen Mechanik an MZP 4

	Modell 1	Modell 2
$n_{\text{Probanden}}$	180	180
LogLike.	−5211.40	−5203.53
Deviance	10 422.83	10 407.05
AIC	10 471	10 469
BIC	10 547	10 568
CAIC	10 571	10 599
Geschätzte Parameter	24	31
Dimension 1		
n_{Items}	79	31
EAP-Reliabilität	0.864	0.843
wMNSQ-Infit	[0.830 ; 1.204*]	[0.828 ; 1.271]
wMNSQ-Outfit	[0.671*; 1.378*]	[0.789 ; 1.371*]
Varianz	0.956	1.043
Dimension 2		
n_{Items}		39
EAP-Reliabilität		0.847
wMNSQ-Infit		[0.740*; 1.199]
wMNSQ-Outfit		[0.882 ; 1.319*]
Varianz		1.083
Dimension 3		
n_{Items}		9
EAP-Reliabilität		0.791
wMNSQ-Infit		[0.898 ; 1.054]
wMNSQ-Outfit		[0.820 ; 1.069]
Varianz		0.605

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität.
Modell 1 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Rasch-Modell (1pl, 3-Dim): Dimension 1 = Fachwissen, Dimension 2 = Modellierungsfähigkeit, Dimension 3 = Items der Technischen Mechanik 2 (TM_4).
* = Ein Item liegt außerhalb der Grenzen.

Tabelle 6.16
Kennwerte IRT-Modelle (2pl und 3pl) - Testinstrument der Technischen Mechanik an MZP 4

	Modell 3	Modell 4
$n_{\text{Probanden}}$	180	180
LogLike.	−5114.71	−5157.88
Deviance	10 229.41	10 315.758
AIC	10 433	10 522
BIC	10 759	10 851
CAIC	10 861	10 954
Geschätzte Parameter	102	103
n_{Items}	79	79
EAP-Reliabilität	0.871	0.848
wMNSQ-Infit	[0.865 ; 1.237*]	[0.701*; 1.199]
wMNSQ-Outfit	[0.765*; 1.307*]	[0.859 ; 1.275*]
Varianz	1.000	3.011

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 3 = Birnbaum-Modell (2pl, 1-Dim);
Modell 4 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

Wie bei den beiden vorhergehenden MZP besitzt Modell 3 eine bessere Passung zu den vorhanden Daten. Modell 3 erreicht geringere Informationskriterien und eine höhere EAP-Reliabilität als das vierte Modell. Die Aussage der beiden Kriterien wird auch durch das dritte Kriterium, den χ^2-*Anpassungstest* unterstützt, der keine signifikante Verbesserung für Modell 4 zeigt ($\chi^2(1) = -86.343$, $p > .999$).

Da das 1pl 3-dimensionale Modell (Modell 2) eine bessere Passung zu den vorhanden Daten als das 1-dimensionale Modell (Modell 1) zeigt, werden auch die Erweiterungen des 3-dimensionalen Modells mit dem 2pl- (Modell 5) und 3pl-Modell (Modell 6) berechnet:

- Modell 5 = Birnbaum-Modell (2pl, 3-Dim)
- Modell 6 = Birnbaum-Modell (3pl, 3-Dim).

Die Kennwerte für die beiden 3-dimensionalen Birnbaum-Modelle sind in Tab. 6.17 aufgelistet.

Modell 5 hat niedrigere Werte für alle drei Informationskriterien. Die EAP-Reliabilitäten sind bei Modell 5 für alle Dimensionen höher als bei Modell 6. Nach diesen beiden Kriterien ist das 2pl-Birnbaum-Modell (Modell 5) dem 3pl-Birnbaum-Modell (Modell 6) vorzuziehen. Dies wird auch durch den χ^2-*Anpassungstest* ($\chi^2(3) = -91.743$, $p > .999$) bestätigt, der kein signifikantes Ergebnis liefert und somit das Modell mit der geringeren Anzahl an frei geschätzten Parametern (Modell 5) bevorzugt.

Welches Modell insgesamt die beste Passung für die Daten von MZP 4 aufweist, soll abschließend geklärt werden. Dazu sind die Kennwerte der 3-dimensionalen 1pl- und 2pl-Modelle sowie des 1-dimensionalen 2pl-Modells nochmals in Tab. 6.18 zusammengestellt.

Im Vergleich der drei Modelle zeigt Modell 2 die geringste Passung zu den Daten. Die Informationskriterien liefern keine eindeutigen Aussagen. Die EAP-Reliabilitäten sind für die ersten beiden Dimensionen niedriger, nur die dritte Dimension hat einen höheren Wert als die entsprechende Dimension in Modell 5. Signifikante Ergebnisse zeigt der χ^2-*Anpassungstest* von Modell 2 gegenüber Modell 3 ($\chi^2(71) = 177.644$, $p < .001$) und Modell 5 ($\chi^2(76) = 187.563$, $p < .001$). Somit ist das

Tabelle 6.17
Kennwerte IRT-Modelle (2pl und 3pl, 3-dimensional) Testinstrument der Technischen Mechanik an MZP 4

	Modell 5	Modell 6
$n_{\text{Probanden}}$	180	180
LogLike.	−5109.75	−5155.62
Deviance	10 219.49	10 311.23
AIC	10 433	10 531
BIC	10 775	10 882
CAIC	10 882	10 992
Geschätzte Parameter	107	110
Dimension 1		
n_{Items}	31	31
EAP-Reliabilität	0.846	0.835
wMNSQ-Infit	[0.926 ; 1.249*]	[0.749*; 1.125]
wMNSQ-Outfit	[0.754*; 1.195]	[0.768*; 1.241*]
Varianz	1	2.821
Dimension 2		
n_{Items}	39	39
EAP-Reliabilität	0.852	0.827
wMNSQ-Infit	[0.865 ; 1.243*]	[0.814 ; 1.203*]
wMNSQ-Outfit	[0.820 ; 1.141]	[0.759*; 1.349*]
Varianz	1	3.026
Dimension 3		
n_{Items}		9
EAP-Reliabilität	0.701	0.755
wMNSQ-Infit	[0.994 ; 1.027]	[0.810 ; 1.057]
wMNSQ-Outfit	[0.890 ; 1.021]	[0.712*; 1.023]
Varianz	1	2.181

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität.
Modell 5 = Birnbaum-Modell (2pl, 3-Dim);
Modell 6 = Birnbaum-Modell (3pl, 3-Dim): Rateparameter auf .25 festgesetzt
Dimension 1 = Fachwissen, Dimension 2 = Modellierungsfähigkeit, Dimension 3 = Items der Technischen Mechanik 2.
* = Ein Item liegt außerhalb der Grenzen.

Tabelle 6.18
Kennwerte IRT-Modelle (1pl, 3-dimensional sowie 2pl, 1- und 3-dimensional) - Testinstrument der Technischen Mechanik an MZP 4

	Modell 2	Modell 3	Modell 5
AIC	10 469	10 433	10 433
BIC	10 568	10 759	10 755
CAIC	10 599	10 861	10 882
Geschätzte Parameter	31	102	107
Dimension 1			
EAP-Reliabilität	0.843	0.871	0.846
Dimension 2			
EAP-Reliabilität	0.847		0.852
Dimension 3			
EAP-Reliabilität	0.791		0.701

Anmerkungen. MZP = Messzeitpunkt; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; EAP-Reliabilität = Expected a Posteriori-Reliabilität.
Modell 2 = Rasch-Modell (1pl, 3-Dim);
Modell 3 = Birnbaum-Modell (2pl, 1-Dim);
Modell 5 = Birnbaum-Modell (2pl, 3-Dim);
Dimension 1 = Fachwissen, Dimension 2 = Modellierungsfähigkeit, Dimension 3 = Items der Technischen Mechanik 2.

3-dimensionale 1pl-Rasch-Modell den beiden 2pl-Birnbaum-Modellen bei den vorliegenden Daten unterlegen.

Der Vergleich der beiden 2pl-Birnbaum-Modelle (1- und 3-dimensional) zeigt keine eindeutigen Ergebnisse. Bei den Informationskriterien sind die Werte des AIC identisch, die des BIC bei Modell 5 und die des CAIC bei Modell 3 etwas niedriger. Die EAP-Reliabilitäten bevorzugen das Modell 3, was auch durch den χ^2-*Anpassungstest* bestätigt werden kann, der kein signifikantes Ergebnis ($\chi^2(5) = 9.919$, $p = .078$) zeigt. Da Modell 5 mehr frei geschätzte Parameter hat (Sparsamkeitskriterium) und keine eindeutig bessere Passung aufweist, ist Modell 3 (2pl, 1-dimensional) vorzuziehen.

Wie aus Tab. 6.19 hervorgeht, beträgt die durchschnittliche Itemschwierigkeit $\beta_M = 0.55$ und die Personenfähigkeit $\theta_M = 0.87$. Im Vergleich zu MZP 3 hat also die mittlere Itemschwierigkeit zugenommen ($\beta_M = 0.37$), während die Personenfähigkeit sogar leicht abgenommen hat ($\theta_M = 0.89$). Die Daten deuten also darauf hin, dass im zweiten Fachsemester kein Leistungszuwachs (Technische Mechanik) stattgefunden hat. Eine weitere Interpretationsmöglichkeit wäre, dass die verwendeten Items die gelehrten Inhalte der Technischen Mechanik 2 nicht ausreichend erfassen.

Die Wright Map der Testinstrumente der Technischen Mechanik an MZP 4 ist in Abb. 6.8 dargestellt. Im Vergleich zur Wright Map von MZP 3 sind die Unterschiede marginal (siehe Abb. 6.7).

Die IRT-Skalierung der Testinstrumente ist damit für alle drei MZP abgeschlossen. Das 2pl-Birnbaum-Modell (1-dimensional) zeigt für alle MZP die beste Passung für den vorliegenden Datensatz. Über die MZP ist zu beobachten, dass die Werte der Iteminformationskriterien abnehmen und die EAP-Reliabilitäten zunehmen. Dies könnte zum einen mit einer besseren Passung der Daten an den jeweiligen MZP zusammenhängen. Der Hauptgrund liegt jedoch darin, dass durch die Verwendung von Ankeritems weniger Parameter geschätzt werden müssen und sich dadurch entsprechende Ausprägungen ergeben.

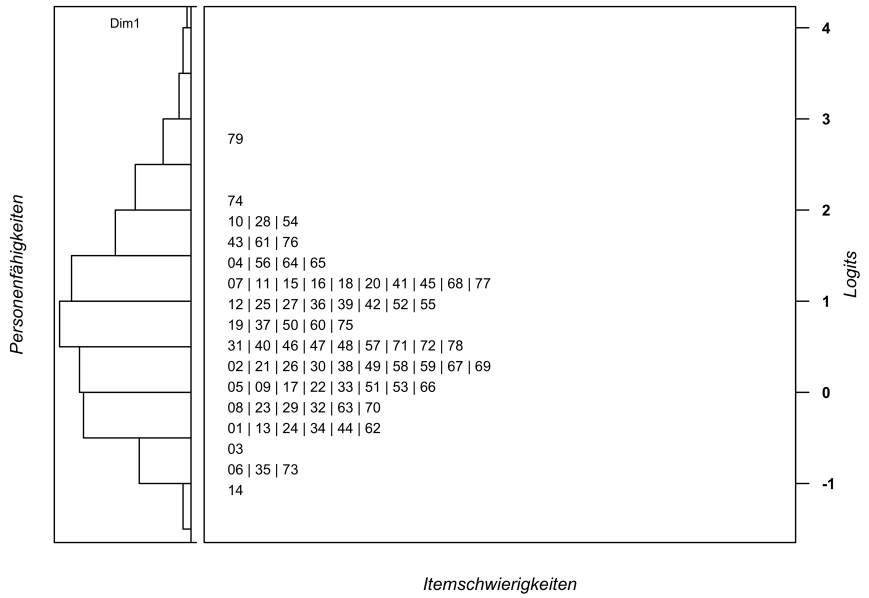
Im Folgenden wird die IRT-Skalierung der Testinstrumente der Rechenfähigkeit vorgenommen.

Tabelle 6.19
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 2pl) - Testinstrument der Technischen Mechanik an MZP 4

	MZP 4
Itemschwierigkeiten	
n_{Items}	79
β_M	0.55
β_{Min}	−1.18
β_{Max}	2.76
Personenfähigkeiten	
$n_{\text{Probanden}}$	180
θ_M	0.87
θ_{Min}	−3.09
θ_{Max}	4.48

Anmerkungen. MZP = Messzeitpunkt; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Abbildung 6.8
Wright Map - Testinstrument der Technischen Mechanik an MZP 4



6.2.3.2 Rechenfähigkeit (FUNDAMENT)

Die Überprüfung der Skalierbarkeit in Kap. 6.2.1 hat ergeben, dass die Testinstrumente zur Erfassung der Rechenfähigkeit für die beiden Studien getrennt voneinander skaliert werden müssen.

Zunächst erfolgt die Skalierung der in FUNDAMENT verwendeten Testinstrumente. Analog zur Skalierung der Testinstrumente der Technischen Mechanik werden die Testinstrumente der Rechenfähigkeit in »R« (R Core Team, 2023) mit dem Paket ›TAM‹ (Robitzsch et al., 2022) und den Funktionen ›tam.mml‹, ›tam.mml.2pl‹ bzw. ›tam.mml.3pl‹ skaliert. Die Schätzung der Personenfähigkeiten erfolgt mit dem *Weighted Maximum Likelihood Estimates*-Schätzer und der entsprechenden Funktion ›tam.wle‹.

Da *a priori* keine Dimensionen der Rechenfähigkeit bekannt sind, werden ausschließlich 1-dimensionale Modelle berechnet. So werden für jeden MZP drei verschiedene Modelle skaliert:

- Modell 1 = Rasch-Modell (1pl, 1-Dim)
- Modell 2 = Birnbaum-Modell (2pl, 1-Dim)
- Modell 3 = Birnbaum-Modell (3pl, 1-Dim).

Bei den MZP 3 und 4 werden für die Ankeritems die Itemschwierigkeiten des vorhergehenden MZP fixiert. Bei Modell 3 erfolgt die Fixierung des Rateparameters für jedes Item auf $\gamma_j = .25$.

Die Überprüfung von Item-DIF erfolgt analog zu den Testinstrumenten der Technischen Mechanik. An keinem der drei MZP weist ein Item für alle vier verwendeten Verfahren Auffälligkeiten auf, sodass alle für die Skalierung verwendeten Items im Datensatz verbleiben.

MZP 2 In FUNDAMENT wurden 20 Items des Testinstruments der Rechenfähigkeit am zweiten MZP eingesetzt (siehe Tab. 6.6). Wie oben beschrieben, werden mit diesen Daten drei Modelle berechnet. Die entsprechenden Kennwerte sind in Tab. 6.26 aufgelistet.

Wie bei den Testinstrumenten der Technischen Mechanik wird der Modellvergleich anhand von drei Kriterien durchgeführt: den Informationskriterien (AIC, BIC und CAIC), den EAP-Reliabilitäten und dem

Tabelle 6.20
Kennwerte IRT-Modelle (1pl, 2pl und 3pl) - Testinstrument der Rechenfähigkeit an MZP 2 - FUNDAMENT

	Modell 1	Modell 2	Modell 3
$n_{\text{Probanden}}$	274	274	274
LogLike.	−3240.98	−3214.80	−3208.97
Deviance	6481.97	6429.61	6417.95
AIC	6524	6510	6500
BIC	6600	6654	6648
CAIC	6621	6694	6689
Geschätzte Par.	21	40	41
n_{Items}	20	20	20
EAP-Reliabilität	0.719	0.739	0.718
wMNSQ-Infit	[0.860; 1.084]	[0.975; 1.009]	[0.766*; 1.021]
wMNSQ-Outfit	[0.821; 1.187]	[0.958; 1.075]	[0.761*; 1.247*]
Varianz	0.950	1.000	1.425

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; Geschätzte Par. = Geschätzte Parameter; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 2 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Birnbaum-Modell (2pl, 1-Dim);
Modell 3 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

χ^2-*Anpassungstest*. Das Modell mit den niedrigsten Werten für die Informationskriterien ist das Modell mit der besten Passung zu den vorliegenden Daten. Die Informationskriterien liefern kein klares Ergebnis. Während Modell 3 den besten Wert des AIC aufweist, sind die des BIC und CAIC für Modell 1 am niedrigsten. Die EAP-Reliabilität ist bei Modell 2 am höchsten, Modell 1 hat einen geringfügig niedrigeren Wert als das dritte Modell, wobei dieser Unterschied im Bereich der Rundungsfehler liegt. Der χ^2-*Anpassungstest* zeigt signifikante Ergebnisse für Modell 1 gegenüber Modell 2 ($\chi^2(19) = 52.363$, $p < .001$) und auch Modell 3 ($\chi^2(20) = 64.020$, $p < .001$). Die Birnbaum-Modelle (2pl und 3pl) sind daher dem Rasch-Modell (1pl) vorzuziehen. Ein signifikantes Ergebnis ($\chi^2(1) = 11.657$, $p < .001$) ergibt sich auch bei dem Vergleich von Modell 2 und Modell 3, sodass Modell 3 vorzuziehen wäre. Da jedoch die EAP-Reliabilität geringer ist, wird davon abgesehen und ebenfalls aus Gründen der Sparsamkeit – auch in Analogie zu den Testinstrumenten der Technischen Mechanik – das Modell 2 weiter verwendet.

Tabelle 6.21
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 2pl) - Testinstrument der Rechenfähigkeit an MZP 2 - FUNDAMENT

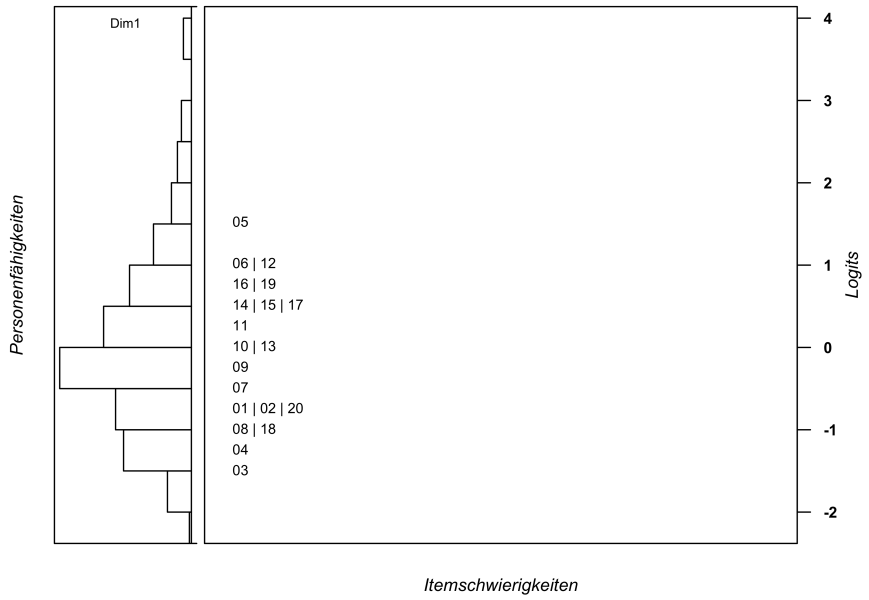
	MZP 2 RF
Itemschwierigkeiten	
n_{Items}	20
β_M	−0.05
β_{Min}	−1.62
β_{Max}	1.6
Personenfähigkeiten	
$n_{\text{Probanden}}$	274
θ_M	0.04
θ_{Min}	−3.05
θ_{Max}	5.11

Anmerkungen. MZP = Messzeitpunkt; RF = Rechenfähigkeit; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

An diesem MZP liegt die mittlere Itemschwierigkeit bei $\beta_M = -0.05$ und die Personenfähigkeit bei $\theta_M = 0.04$. Weitere Item- und Probandenkennwerte sind in Tab. 6.21 aufgeführt.

Die dazugehörige Wright Map des Testinstruments der Rechenfähigkeit an MZP 2 (FUNDAMENT) ist in Abb. 6.9 dargestellt.

Abbildung 6.9
Wright Map - Testinstrument der Rechenfähigkeit an MZP 2 - FUNDAMENT



MZP 3 Am dritten MZP erfolgt eine Verankerung der bereits an MZP 2 eingesetzten Items (Ankeritems). Die geschätzten Itemschwierigkeiten des 2pl-Birnbaum-Modells (MZP 2) werden für die entsprechenden Items an MZP 3 fixiert. Von den 12 verwendeten Items sind zehn Ankeritems, entsprechend sind weniger Parameter zu schätzen. Diese Angabe steht im Widerspruch zu Tab. 5.3, da dort angegeben wird, dass alle eingesetzten Items auch Ankeritems sind. Dies ist theoretisch richtig, allerdings wurden Ankeritems bei der Modellpassung (siehe Kap. 6.2.2) aus dem Datensatz an MZP 2 ausgeschlossen, sodass diese Items erst an MZP 3 verwendet wurden und somit keine Anker mehr darstellen.

Die Kennwerte des 1pl-Rasch-Modells (Modell 1) und der beiden

Birnbaum-Modelle (2pl- [Modell 2] und 3pl [Modell 3]) sind in Tab. 6.22 aufgelistet.

Tabelle 6.22
Kennwerte IRT-Modelle (1pl, 2pl und 3pl) - Testinstrument der Rechenfähigkeit an MZP 3 - FUNDAMENT

	Modell 1	Modell 2	Modell 3
$n_{\text{Probanden}}$	114	114	114
LogLike.	−816.87	−802.91	−805.73
Deviance	1633.74	1605.82	1611.47
AIC	1642	1636	1643
BIC	1653	1677	1687
CAIC	1657	1692	1703
Geschätzte Par.	4	15	16
n_{Items}	12	12	12
EAP-Reliabilität	0.697	0.713	0.715
wMNSQ-Infit	[0.850 ; 1.124]	[0.902; 1.055]	[0.886; 1.019]
wMNSQ-Outfit	[0.768*; 1.178]	[0.837; 1.165]	[0.878; 1.112]
Varianz	1.168	1.000	0.956

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; Geschätzte Par. = Geschätzte Parameter; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 2 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Birnbaum-Modell (2pl, 1-Dim);
Modell 3 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

Die Anzahl der geschätzten Parameter wird durch die Festlegung der Itemschwierigkeiten der Ankeritems deutlich reduziert. Während an MZP 2 noch 21 Parameter geschätzt werden mussten, sind es an MZP 3 nur noch 4 Parameter. Bei den beiden anderen Modellen hat sich die Anzahl der zu schätzenden Parameter mehr als halbiert (von 40 auf 15 Parameter). Dementsprechend haben sich auch die Werte der Informationskriterien deutlich verringert. Modell 2 hat die niedrigsten BIC und CAIC, das AIC ist bei Modell 2 am niedrigsten, während

Modell 3 bei allen die höchsten und damit schlechtesten Werte erzielt. Die EAP-Reliabilität ist dagegen bei Modell 3 am höchsten, wohingegen das erste Modell die niedrigste EAP-Reliabilität aufweist. Der χ^2-*Anpassungstest* zeigt eine signifikant bessere Passung der Daten für die Modelle 2 ($\chi^2(11) = 27.920$, $p = .003$) und 3 ($\chi^2(12) = 22.269$, $p = .035$) im Vergleich zu Modell 1. Modell 3 zeigt jedoch keine bessere Passung als Modell 2 ($\chi^2(19) = -5 - 651$, $p > .999$). Somit wird, wie bei dem vorherigen MZP, Modell 2 (2pl-Birnbaum-Modell) als am besten geeignet für die Daten angesehen und dementsprechend weiter verwendet.

Die mittlere Itemschwierigkeit liegt an diesem MZP bei $\beta_M = 0.81$ und die Personenfähigkeit bei $\theta_M = 1.34$ (siehe Tab. 6.23). Beide Werte haben somit deutlich gegenüber dem zweiten MZP zugenommen ($\beta_M = -0.05$; $\theta_M = 0.04$).

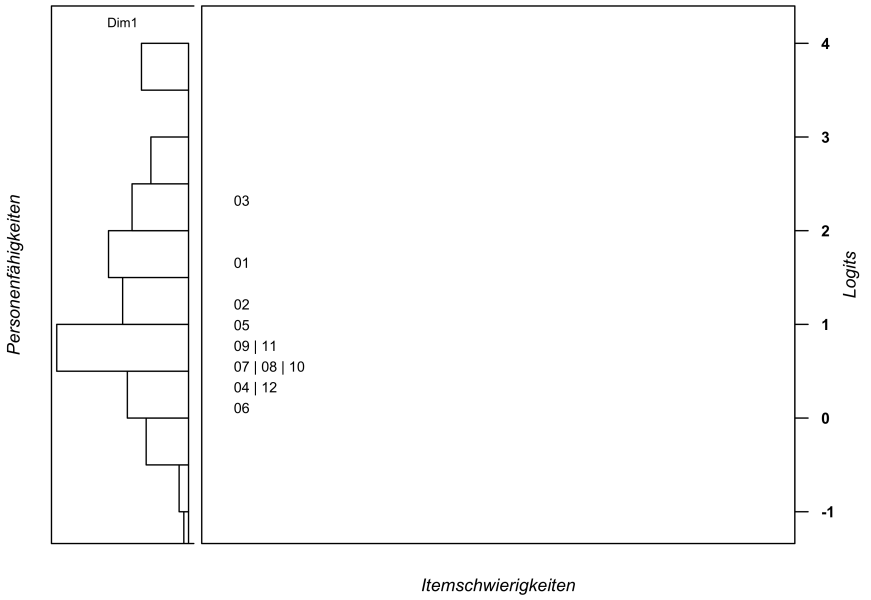
Tabelle 6.23
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 2pl) - Testinstrument der Rechenfähigkeit an MZP 3 - FUNDAMENT

	MZP 3 RF
Itemschwierigkeiten	
n_{Items}	12
β_M	0.81
β_{Min}	0.12
β_{Max}	2.37
Personenfähigkeiten	
$n_{\text{Probanden}}$	114
θ_M	1.34
θ_{Min}	−1.11
θ_{Max}	3.94

Anmerkungen. MZP = Messzeitpunkt; RF = Rechenfähigkeit; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Dies zeigt sich auch in der Wright Map (Abb. 6.10) zum Testinstrument der Rechenfähigkeit an MZP 3 (FUNDAMENT). Sowohl die Itemschwierigkeiten, als auch die Personenfähigkeiten sind auf der Logit-Skala in positive Richtung verschoben.

Abbildung 6.10
Wright Map - Testinstrument der Rechenfähigkeit an MZP 3 - FUNDAMENT



MZP 4 Wie zuvor werden die Itemschwierigkeiten derjenigen Items verankert, die auch am folgenden MZP verwendet werden. Für die Ankeritems werden die Itemschwierigkeiten des 2pl-Birnbaum-Modells aus MZP 3 übernommen. Zehn der insgesamt 12 Items sind Ankeritems (siehe Tab. 6.6). Auch wenn diese Zahlen mit MZP 3 identisch sind, handelt es sich bei den zehn Ankeritems nicht bei allen Items um dieselben, die bereits von MZP 2 nach MZP 3 verankert wurden. Die Kennwerte der mit den Daten skalierten Modelle finden sich in Tab. 6.24.

Da die gleiche Anzahl von Items wie an MZP 3 verwendet wird und auch die Anzahl der Ankeritems gleich ist, ist auch die Anzahl der zu schätzenden Parameter gleich. Modell 1 hat für alle drei Informationskriterien die höchsten und damit schlechtesten Werte. Während das AIC bei Modell 3 um einen Punkt niedriger und damit besser ist als bei Modell 2, liegen die Werte des BIC und des CAIC um zwei bzw. drei Punkte höher als beim zweiten Modell. Alle EAP-Reliabilitäten liegen in einem akzeptablen Bereich, wobei Modell 3 den besten Wert vor Modell 1 und Modell 2 aufweist, der Unterschied zwischen diesen beiden

Tabelle 6.24
Kennwerte IRT-Modelle (1pl, 2pl und 3pl) - Testinstrument der Rechenfähigkeit an MZP 4 - FUNDAMENT

	Modell 1	Modell 2	Modell 3
$n_{\text{Probanden}}$	105	105	105
LogLike.	−787.83	−737.4	−736.09
Deviance	1575.67	1474.81	1472.17
AIC	1584	1505	1504
BIC	1594	1545	1547
CAIC	1598	1560	1563
Geschätzte Par.	4	15	16
n_{Items}	12	12	12
EAP-Reliabilität	0.741	0.739	0.780
wMNSQ-Infit	[0.875; 1.641*]	[0.942; 1.218*]	[0.956; 1.230*]
wMNSQ-Outfit	[0.835; 1.908*]	[0.858; 1.337*]	[0.956; 1.092]
Varianz	1.485	1.000	0.746

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; Geschätzte Par. = Geschätzte Parameter; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 2 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Birnbaum-Modell (2pl, 1-Dim);
Modell 3 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

jedoch sehr gering ist. Der χ^2-*Anpassungstest* zeigt signifikante Verbesserungen für die beiden Modelle mit mehr geschätzten Parametern, Modell 2 ($\chi^2(11) = 100.864$, $p < .001$) und Modell 3 ($\chi^2(12) = 103.499$, $p < .001$), gegenüber Modell 1. Modell 3 hingegen stellt keine signifikante Verbesserung gegenüber Modell 2 dar ($\chi^2(1) = 2.635$, $p = .105$).

Damit zeigt Modell 2 (2pl-Birnbaum-Modell), wie bereits bei den beiden vorangegangen MZP, die beste Passung zu den vorliegenden Daten. Die zugehörigen Itemschwierigkeiten und Personenfähigkeiten sind in Tab. 6.25 aufgelistet. Die mittlere Itemschwierigkeit beträgt $\beta_M = 0.92$ und ist damit um 0.11 höher als bei MZP 2. Die mittlere Personenfähigkeit beträgt $\theta_M = 1.32$ und ist damit um 0.02 niedriger als bei dem vorherigen MZP.

Tabelle 6.25
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 2pl) - Testinstrument der Rechenfähigkeit an MZP 4 - FUNDAMENT

	MZP 4 RF
Itemschwierigkeiten	
n_{Items}	12
β_M	0.92
β_{Min}	0.12
β_{Max}	2.37
Personenfähigkeiten	
$n_{\text{Probanden}}$	105
θ_M	1.32
θ_{Min}	−2.36
θ_{Max}	4.29

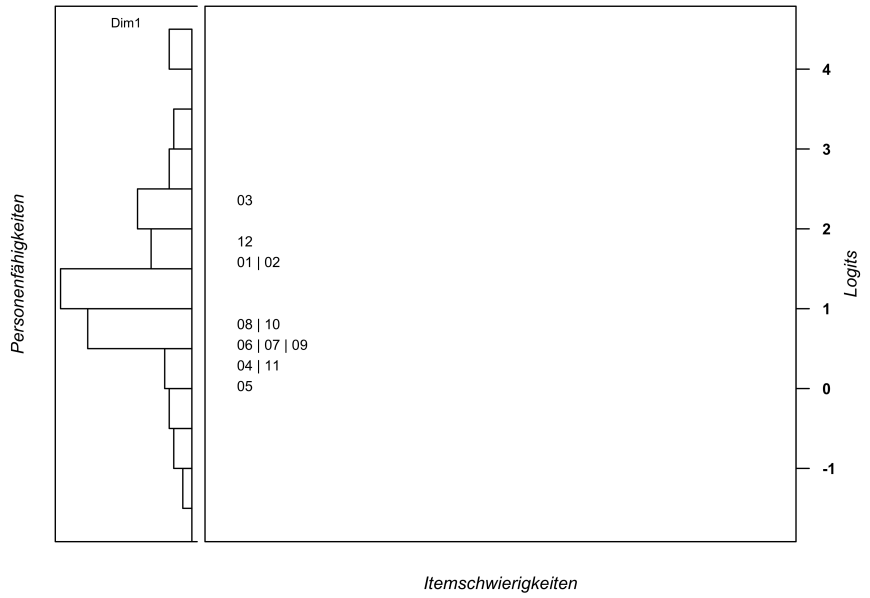
Anmerkungen. MZP = Messzeitpunkt; RF = Rechenfähigkeit; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Die Itemschwierigkeit und das Histogramm der Personenfähigkeiten sind in Abb. 6.11 als Wright Map zum Testinstrument der Rechenfähigkeit an MZP 4 (FUNDAMENT) visualisiert.

Die Testinstrumente der Rechenfähigkeit aller drei MZP der Studie FUNDAMENT sind somit nach dem 2pl-Birnbaum-Modell IRT skaliert. Das 2pl-Birnbaum-Modell zeigt für die drei MZP die beste Passung für

den vorliegenden Datensatz.

Abbildung 6.11
Wright Map - Testinstrument der Rechenfähigkeit an MZP 4 - FUNDAMENT



Nachfolgend wird das Testinstrument der Rechenfähigkeit der Studie ALSTER IRT skaliert.

6.2.3.3 Rechenfähigkeit (ALSTER)

Bereits in Kap. 6.2.1 wurde bei der Überprüfung der Skalierbarkeit festgestellt, dass eine separate Skalierung der Testinstrumente der Rechenfähigkeit der Studie ALSTER notwendig ist. Die Skalierung erfolgt in Analogie zu den Testinstrumenten der Technischen Mechanik sowie der Rechenfähigkeit von FUNDAMENT. Erneut werden drei Modelle für jeden MZP skaliert:

- Modell 1 = Rasch-Modell (1pl, 1-Dim)
- Modell 2 = Birnbaum-Modell (2pl, 1-Dim)
- Modell 3 = Birnbaum-Modell (3pl, 1-Dim).

Bei der Skalierung der beiden MZP – für den vierten MZP liegen aufgrund von Kodierungsproblemen keine verwertbaren Daten vor – werden die Itemschwierigkeiten des zweiten MZP für Ankeritems an MZP 3 fixiert. Der Rateparameter des dritten Modells wird wiederum für jedes Item auf $\gamma_j = .25$ festgesetzt.

Die verwendeten Items werden ebenfalls mit den oben genannten Verfahren auf Item-DIF getestet. Auch hier zeigt keines der Items Auffälligkeiten in allen vier Verfahren, folglich bleiben alle Items im Datensatz.

MZP 2 14 Items wurden am zweiten MZP zur Erfassung der Rechenfähigkeit in ALSTER eingesetzt (siehe Tab. 6.6). Die Kennwerte der drei skalierten Modelle finden sich in Tab. 6.26.

Im Gegensatz zu allen vorherigen Skalierungen ist das 1pl-Rasch-Modell das Modell mit der besten Passung zu den vorliegenden Daten. Modell 1 hat für alle drei Informationskriterien die niedrigsten Werte, während das dritte Modell für alle die höchsten Werte und somit die schlechteste Passung aufweist. Alle EAP-Reliabilitäten sind als akzeptabel zu bewerten, mit einem Wert von 0.787 erreicht Modell 1 den zweitbesten Wert, wogegen Modell 2 die beste Passung erreicht. Der χ^2-*Anpassungstest* liefert keine signifikanten Ergebnisse. Modell 3 stellt keine Verbesserung von Modell 2 dar ($\chi^2(1) = -104.767$, $p > .999$), aber beide Modelle verbessern auch nicht signifikant das erste Modell (Modell 2: $\chi^2(13) = 14.113$, $p = .366$; Modell 3: $\chi^2(14) = -90.654$, $p > .999$).

Das Modell 1 passt demnach am besten zu den vorliegenden Daten, sodass das 1pl-Rasch-Modell weiter verwendet wird. Die mit dem Modell berechneten Itemschwierigkeiten und Personenfähigkeiten sind in Tab. 6.27 aufgeführt. Die mittlere Itemschwierigkeit beträgt $\beta_M = -0.24$ und die mittlere Personenfähigkeit $\theta_M = -0.02$.

Die zu den Daten gehörende Wright Map des Testinstruments der Rechenfähigkeit an MZP 2 (ALSTER) ist in Abb. 6.12 abgebildet.

Tabelle 6.26
Kennwerte IRT-Modelle (1pl, 2pl und 3pl) - Testinstrument der Rechenfähigkeit an MZP 2 - ALSTER

	Modell 1	Modell 2	Modell 3
$n_{\text{Probanden}}$	193	193	193
LogLike.	−1352.38	−1345.33	−1397.71
Deviance	2704.77	2690.65	2795.42
AIC	2735	2747	2853
BIC	2784	2838	2948
CAIC	2799	2866	2977
Geschätzte Par.	15	28	29
n_{Items}	14	14	14
EAP-Reliabilität	0.787	0.794	0.753
wMNSQ-Infit	[0.903; 1.127]	[0.969; 1.023]	[0.726*; 0.992]
wMNSQ-Outfit	[0.825; 1.198]	[0.869; 1.062]	[0.720*; 1.010]
Varianz	2.169	1.000	1.643

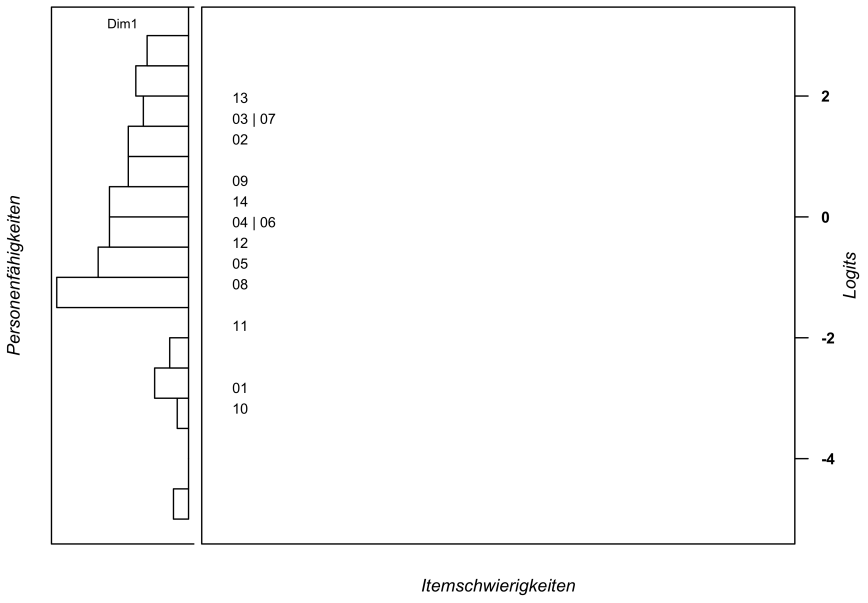
Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; Geschätzte Par. = Geschätzte Parameter; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 2 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Birnbaum-Modell (2pl, 1-Dim);
Modell 3 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

Tabelle 6.27
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 1pl) - Testinstrument der Rechenfähigkeit an MZP 2 - ALSTER

	MZP 2 RF
Itemschwierigkeiten	
n_{Items}	14
β_M	−0.24
β_{Min}	−3.1
β_{Max}	2.06
Personenfähigkeiten	
$n_{\text{Probanden}}$	193
θ_M	−0.02
θ_{Min}	−4.83
θ_{Max}	4.06

Anmerkungen. MZP = Messzeitpunkt; RF = Rechenfähigkeit; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Abbildung 6.12
Wright Map - Testinstrument der Rechenfähigkeit an MZP 2 - ALSTER



MZP 3 Der dritte MZP besteht aus 10 Items, von denen 6 Ankeritems sind (siehe Tab. 6.6). Bekanntlich sind die geschätzten Itemschwierigkeiten des vorhergehenden MZP (1pl-Rasch-Modell) für MZP 3 fixiert. Die Kennwerte der Modelle mit fixierten Itemschwierigkeiten finden sich in Tab. 6.28.

Tabelle 6.28
Kennwerte IRT-Modelle (1pl, 2pl und 3pl) - Testinstrument der Rechenfähigkeit an MZP 3 - ALSTER

	Modell 1	Modell 2	Modell 3
$n_{\text{Probanden}}$	184	184	184
LogLike.	−981.82	−969.28	−998.5
Deviance	1963.63	1938.56	1997.15
AIC	1976	1969	2029
BIC	1995	2017	2081
CAIC	2001	2032	2097
Geschätzte Par.	6	15	16
n_{Items}	10	10	10
EAP-Reliabilität	0.797	0.801	0.704
wMNSQ-Infit	[0.909; 1.113]	[0.908; 1.147]	[0.678*; 1.010]
wMNSQ-Outfit	[0.843; 1.392*]	[0.841; 1.302*]	[0.672*; 1.447*
Varianz	3.253	1.000	2.129]

Anmerkungen. MZP = Messzeitpunkt; $n_{\text{Probanden}}$ = Anzahl der Probanden; LogLike. = statistischer Wert der logarithmischen Likelihood-Funktion; Deviance = Anpassungsgüte des IRT-Modells; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion; Geschätzte Par. = Geschätzte Parameter; n_{Items} = Anzahl der Items; EAP-Reliabilität = Expected a Posteriori-Reliabilität; wMNSQ-Infit/Outfit = Weighted-Mean-Square-Infit/Outfit.
Modell 2 = Rasch-Modell (1pl, 1-Dim);
Modell 2 = Birnbaum-Modell (2pl, 1-Dim);
Modell 3 = Birnbaum-Modell (3pl, 1-Dim): Rateparameter auf .25 festgesetzt.
* = Ein Item liegt außerhalb der Grenzen.

Die Modellprüfung der Informationskriterien weist Modell 3 als das Modell mit der schlechtesten Passung der drei Modelle aus. Dies bestätigt auch der χ^2-*Anpassungstest*, der weder für Modell 1 ($\chi^2(10) = -33.515$, $p > .999$), noch für Modell 2 ($\chi^2(1) = -58.581$, $p > .999$) signifikante Ergebnisse zeigt. Modell 1 zeigt die niedrigsten Werte für das BIC

und das CAIC, während Modell 2 das niedrigste AIC aufweist. Die EAP-Reliabilität ist bei Modell 2 am höchsten und kann sogar als gut eingestuft werden, während Modell 1 einen etwas niedrigeren Wert aufweist und als akzeptabel zu bewerten ist. Die bessere Passung des zweiten Modells wird auch durch das signifikante Ergebnis des χ^2-*Anpassungstest* belegt ($\chi^2(9) = 25.066$, $p = .003$).

Da sowohl das AIC als auch die EAP-Reliabilität und der χ^2-*Anpassungstest* dem Modell 2 die beste Passung der Daten zuweisen, wäre die Wahl des 2pl-Birnbaum-Modell für MZP 3 eigentlich angemessen. Davon wird jedoch abgesehen, da der zweite MZP 1pl-Rasch skaliert wurde. Dementsprechend wird analog auch MZP 3 nach Modell 1 skaliert. Diese Entscheidung ist auch vertretbar, da das BIC und das CAIC die geringsten Werte für Modell 1 aufweisen und somit die beste Modellpassung darstellen. Die EAP-Reliabilität ist zwar geringer als bei MZP 2, aber der Unterschied ist mit .007 sehr gering. Daher wird im Folgenden für den dritten MZP die Skalierung des 1pl-Rasch-Modell verwendet. Die resultierende mittlere Itemschwierigkeit beträgt $\beta_M = 0.34$ (siehe Tab. 6.29), damit ist ein deutlicher Anstieg der Itemschwierigkeit gegenüber dem vorherigen MZP zu erkennen ($\beta_M = -0.24$). Während die mittlere Personenfähigkeit mit $\theta_M = 0.07$ nur um 0.09 zunimmt.

Die aus der 1pl-Rasch-Skalierung resultierende Wright Map (Testinstrument der Rechenfähigkeit MZP 3 ALSTER) findet sich in Abb. 6.13. Zu beachten ist, dass sich die Skalierung der Y-Achse (Logits) gegenüber des vorherigen MZP geändert hat.

Die IRT-Skalierung einschließlich der zugehörigen Modellprüfung ist damit abgeschlossen. Es zeigt sich, dass – mit Ausnahme des Testinstruments der Rechenfähigkeit der Studie ALSTER – das 1-dimensionale 2pl-Birnbaum-Modell die beste Passung für den vorliegenden Datensatz liefert.

Mit den nun skalierten Daten werden im folgenden Kapitel Niveaumodelle entwickelt.

Tabelle 6.29
Itemschwierigkeiten/Personenfähigkeiten (IRT-Modell 1pl) - Testinstrument der Rechenfähigkeit an MZP 3 - ALSTER

	MZP 3 RF
Itemschwierigkeiten	
n_{Items}	10
β_M	0.34
β_{Min}	−3.1
β_{Max}	2.84
Personenfähigkeiten	
$n_{\text{Probanden}}$	184
θ_M	0.07
θ_{Min}	−4.28
θ_{Max}	4.42

Anmerkungen. MZP = Messzeitpunkt; RF = Rechenfähigkeit; n = Stichprobengröße; β = Itemschwierigkeit; M = Mittelwert; Min = Minimum; Max = Maximum; θ = Personenfähigkeit.

Abbildung 6.13
Wright Map - Testinstrument der Rechenfähigkeit an MZP 3 - ALSTER

6.3 Entwicklung der Niveaumodelle

Wie in Kap. 4.2.1 erläutert, setzt die Niveaumodellierung die IRT-Skalierung des Datensatzes voraus, und zwar streng genommen die Skalierung als 1pl-Rasch-Modell, da die Parallelität der ICC, d.h. die spezifische Objektivität, gefordert wird (Rauch & Hartig, 2020). In Kap. 4.1.4 hat sich jedoch für die Testinstrumente der Technischen Mechanik und auch für die Testinstrumente der Rechenfähigkeit (FUNDAMENT) gezeigt, dass das 2pl-Birnbaum-Modell eine bessere Passung der Daten liefert. Daher muss auf einen der vier von Wilson (2003) empfohlenen Lösungsansätze zurückgegriffen werden (siehe Kap. 4.2.1). Wie bereits in Kap. 4.2.1 ausgeführt, bietet der zweite Ansatz eine praktikable Lösung. Wilson (2003) bezeichnet diesen Lösungsansatz als »hiding behind the statistical complexity« (Wilson, 2003, S. 16). Dementsprechend wird als hinreichendes Kriterium für die richtige Beantwortung eines Items eine Lösungswahrscheinlichkeit von 80% angenommen. Dieser Wert wird neben Wilson (2003) auch von Beaton und Allen (1992) in ihrer Beschreibung einer Post-hoc-Niveaumodellierung verwendet. Dennoch erscheint der Wert recht hoch, insbesondere im Vergleich zu den in anderen Studien angesetzten Lösungswahrscheinlichkeiten von 65% (siehe Kap. 4.2.1).

Probanden werden einem Kompetenzniveau zugeordnet, wenn sie 50% der Items dieses Niveaus lösen können (OECD, 2004). Diese in PISA (OECD, 2004) verwendete Zuordnung erscheint plausibel und mit der vorliegenden Lösungswahrscheinlichkeit von 80% gut vereinbar. Dies bedeutet, dass ein Proband mit einer Personenfähigkeit von z.B. $\theta = 2.2$, dem nächstliegenden Niveau zugeordnet wird, bei dem die Grenzen der Itemschwierigkeiten (bei $P(u_{ij} = 1) = 80\%$) kleiner oder größer als $\beta = 2.2$ sind. Als Beispiel könnten die Grenzen eines Niveaus bei Itemschwierigkeiten von 2 und 3 liegen, dann würde dieser Proband diesem Niveau zugeordnet werden. Mit $\theta = 2.2$ liegt er jedoch eher am unteren Ende des Niveaus, sodass er die Items mit einer Itemschwierigkeit von $\beta \leq 2.2$ mit einer sehr hohen Lösungswahrscheinlichkeit ($P(u_{ij} = 1) = 80\%$) richtig lösen wird. Bei Items mit einer Itemschwierigkeit von $\beta > 2.2$ wird die Lösungswahrscheinlichkeit entsprechend unter $P(u_{ij} = 1) \leq 80\%$ liegen und damit der Einstufung nach PISA

(OECD, 2004) entsprechen.

Die hohe Lösungswahrscheinlichkeit von 80% führt zwangsläufig zu hohen Itemschwierigkeiten. Die maximale Itemschwierigkeit (bei $P(u_{ij} = 1) = 80\%$) wird auf $\beta < 6$ begrenzt. Die Folge der hohen Lösungswahrscheinlichkeit ist, dass Probanden relativ hohe Personenfähigkeiten aufweisen müssen, um entsprechend höheren Niveaus zugeordnet werden zu können. Diese konservative Auslegung bedeutet auch, dass in den hohen Kompetenzniveaus besser differenziert werden kann als in den niedrigen. Ebenso ist eine schiefe Verteilung zugunsten der niedrigen Kompetenzniveaus zu erwarten.

Da die Daten nach dem 2pl-Birnbaum Modell IRT skaliert wurden, ist es nicht möglich, das zweite Kriterium – Probanden in dem nächstniedrigeren Kompetenzniveau haben eine hinreichend geringe Lösungswahrscheinlichkeit ($P(u_{ij} = 1) \leq 50\%$) – zu erfüllen. Bei einer Lösungswahrscheinlichkeit von 50% können sich die ICC überlagern und zu einer inkonsistenten Rangfolge der Itemschwierigkeiten führen. Aus diesem Grund wird geprüft, ob ein signifikanter Zusammenhang zwischen den Itemschwierigkeiten des verwendeten 2pl-Birnbaum-Modells ($P(u_{ij} = 1) = 80\%$) und dem – nach dem Modellvergleich unpassenden – 1pl-Rasch-Modell ($P(u_{ij} = 1) = 50\%$) besteht. Bei hohen Korrelationen ($|r| \geq 0.5$) kann analog zu Kap. 6.2.1 argumentiert werden, dass die Items die gleiche Rangfolge besitzen. Damit wäre die Verwendung des 2pl-Birnbaum-Modells ($P(u_{ij} = 1) = 80\%$) zur Niveaumodellierung vertretbar (Dammann et al., 2016). Wilson (2003) schreibt die Prüfung der Rangfolge nicht vor und ist daher nur als zusätzliche Absicherung zu verstehen.

Die in ALSTER erhobenen Daten des Testinstruments der Rechenfähigkeit wurden nach dem 1pl-Rasch-Modell skaliert, da dieses Modell die beste Passung zeigt (siehe Kap. 6.2.3.3). Eine Überprüfung der Rangfolge ist daher nicht erforderlich.

Wie bereits bei dem Modellvergleich der IRT-Skalierung (siehe Kap. 6.2.3) erfolgt die Niveaumodellierung zunächst für die Testinstrumente der Technischen Mechanik und anschließend für das Testinstrumente der Rechenfähigkeit getrennt für FUNDAMENT und ALSTER.

6.3.1 Technische Mechanik

Bevor die Niveaumodelle für die Testinstrumente der Technischen Mechanik entwickelt werden können, muss zunächst die Rangfolge der Items überprüft werden (siehe Kap. 6.3). Dieser Schritt ist notwendig, da beim Modellvergleich der IRT-Skalierung das 2pl-Birnbaum-Modell die beste Passung zu den vorliegenden Daten aufweist (siehe Kap. 6.2.3.1).

Die Überprüfung der Rangfolge der Items erfolgt durch die Produkt-Moment-Korrelation nach Pearson und dem resultierenden Pearson Korrelationskoeffizienten (r), der die Itemschwierigkeiten des 2pl-Birnbaum-Modells mit einer Lösungswahrscheinlichkeit von $P(u_{ij} = 1) = 80\%$ mit den Itemschwierigkeiten des 1pl-Rasch-Modells ($P(u_{ij} = 1) = 50\%$) vergleicht. Bei hohen signifikanten Korrelationen ($|r| \geq 0.5$) kann von der gleichen Rangfolge der Items gesprochen werden (siehe Kap. 6.2.1). Die Voraussetzungen für die Produkt-Moment-Korrelation nach Pearson (siehe Kap. 6.2.1) sind alle erfüllt, die Berechnung erfolgt erneut mit der Funktion ›cor.test‹ aus dem ›R‹-Paket ›stats‹ (R Core Team, 2023). In Tab. 6.30 sind die berechneten Pearson Korrelationen für die MZP 2 bis 4 für die Testinstrumente der Technischen Mechanik aufgeführt.

Tabelle 6.30
Pearson Korrelationskoeffizienten der Itemschwierigkeiten der Testinstrumente der Technischen Mechanik (2pl-Birnbaum-Modell [$P(u_{ij} = 1) = 80\%$] und 1pl-Rasch-Modell [$P(u_{ij} = 1) = 50\%$])

	t	df	p	r
MZP 2	11.92	63	<.001	.83
MZP 3	8.78	76	<.001	.71
MZP 4	11.30	74	<.001	.80

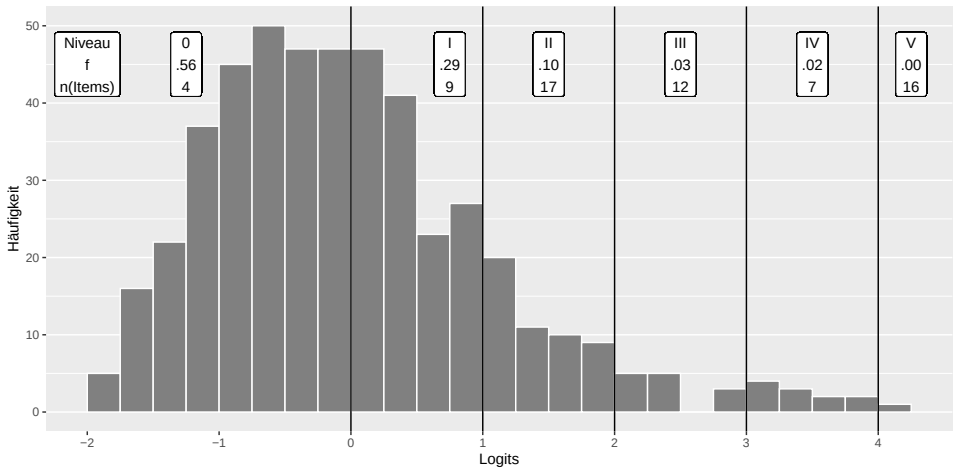
Anmerkungen. t = t-Wert; df = Freiheitsgrade; p = Wahrscheinlichkeit; r = Pearson Korrelationskoeffizient; MZP = Messzeitpunkt.

Es zeigt sich, dass für alle drei MZP hohe signifikante Korrelationen bestehen. Dementsprechend wird für die Itemschwierigkeit der Items des 2pl-Birnbaum-Modells ($P(u_{ij} = 1) = 80\%$) und des 1pl-Rasch-Modells ($P(u_{ij} = 1) = 50\%$) die gleiche Rangfolge angenommen. Somit ist es möglich, die Itemschwierigkeiten des 2pl-Birnbaum-Modells mit einer Lösungswahrscheinlichkeit von 80% für die Niveaumodellierung

zu verwenden.

MZP 2 Für das Testinstrument der Technischen Mechanik an MZP 2 konnten fünf Niveaugrenzen identifiziert werden. Diese Grenzen liegen auf der Logit-Skala bei den Punkten 0, 1, 2, 3 und 4. Aus diesen Grenzen lassen sich sechs Niveaus ableiten, die bezeichnet werden als: 0, I, II, III, IV und V. Diese Bezeichnungen wurden gewählt, da die Niveaus die Itemschwierigkeiten bis zu ihrer Bezeichnung beinhalten, d.h. Niveau 0 beinhaltet alle Items mit einer Itemschwierigkeit (bei $P(u_{ij} = 1) = 80\%$) von $\beta \leq 0$ und Niveau III beinhaltet Itemschwierigkeiten von $2 < \beta \leq 3$. Aufgrund der maximalen Itemschwierigkeit von $\beta < 6$ können 18 Items nicht in dem Niveaumodell berücksichtigt werden, sodass 65 Items am zweiten MZP verbleiben. Das resultierende Niveaumodell ist in Abb. 6.14 dargestellt. Zur Darstellung wird ein Histogramm verwendet, da die Lage der Säulen auf der Achse und die Flächen der Säulen sinnvoll interpretiert werden können (Eid et al., 2017). Die Säulenbreite wurde mit 0.25 Logits festgelegt, multipliziert mit der absoluten Häufigkeit (y-Achse) ergibt sich ein Säulenfläche, die einem Viertel der absoluten Häufigkeit entspricht.

Abbildung 6.14
Niveaumodell des Testinstruments der Technischen Mechanik an MZP 2



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.

Die in Kap. 6.3 gemachten Aussagen bzgl. der Auswirkungen der hohen Lösungswahrscheinlichkeit sind in Abb. 6.14 ersichtlich. Die Verteilung ist rechtsschief und es ist zu beobachten, dass die Kompetenzniveaus I bis V besser diskriminieren können als Niveau 0. Dies ist darauf zurückzuführen, dass die Niveaus I bis V jeweils eine Breite von einem Logit aufweisen, während Niveau 0 zwei Logits umfasst. Theoretisch wäre eine weitere Niveaugrenze bei -1 möglich gewesen, allerdings gibt es keine Items, die entsprechende Itemschwierigkeiten aufweisen. Somit wäre eine inhaltliche Interpretation dieses Niveaus nicht möglich.

Die Verteilung der Probanden ist ebenfalls erwartungsgemäß. Mehr als die Hälfte der Probanden (56%) wurde dem niedrigsten Kompetenzniveau zugeordnet, weitere 29% dem Niveau I. Somit befinden sich 85% aller Probanden in den beiden niedrigsten Niveaus, während in dem höchsten Niveau nur ein Proband vertreten ist. Das bedeutet, dass 56% der Probanden in der Lage sind, die vier Items des Niveaus 0 mit hinreichender Sicherheit zu lösen, wohingegen 29% der Probanden, die Items der Niveaus 0 und I (13 Items) mit hinreichender Sicherheit lösen können.

Inhaltlich ist dies jedoch nachvollziehbar. Schließlich liegt der zweite MZP zu Beginn des ersten Fachsemesters, sodass die Probanden noch keinerlei Veranstaltungen zur Technischen Mechanik gehört haben und die Items des Testinstruments nur mit Vorwissen aus der Schule (Physik), Vorkursen (OV FUNDAMENT) oder dem privaten Umfeld beantworten können.

Insgesamt deuten die Ergebnisse darauf hin, dass das Testinstrument zu schwierig ist. Dies wird auch durch die Anzahl der Items in den verschiedenen Niveaus deutlich. In Niveau 0 gibt es nur 4 Items, wie bereits erwähnt, gibt es keine Items mit geringerer Itemschwierigkeit als $\beta \leq -1$. Das höchste Niveau umfasst dagegen 16 Items, was für diesen MZP nicht ideal ist, da die Probanden offensichtlich (noch) nicht über das nötige Wissen verfügen. Dennoch sollte bedacht werden, dass die dargestellten Itemschwierigkeiten einer Lösungswahrscheinlichkeit von $P(u_{ij} = 1) = 80\%$ entsprechen und damit sehr hoch angesetzt sind.

Die inhaltliche Beschreibung der Niveaus erfolgt anhand der den jeweiligen Niveaus zugeordneten Items. Die Benennung der Themen

orientiert sich dabei am Skript der Vorlesung Technische Mechanik 1 von Schröder (2020). Die in den Items abgefragten Themenbereiche werden nur einmal genannt. Das bedeutet, dass Items mit höherer Komplexität zu bereits genannten Themen der unteren Niveaus in den höheren Niveaus vorkommen können, aber nicht noch einmal explizit genannt werden. Die Probanden der höheren Niveaus verfügen auch über das Wissen der niedrigeren Kompetenzniveaus.

- Niveau 0 ($\beta \leq 0$)
  Probanden dieses Niveaus sind in der Lage, die Einheit der Kraft zu reproduzieren und bei zentralen Kräftesystemen Resultierende und das Gleichgewicht zu bestimmen sowie Kräfte zu zerlegen.

- Niveau I ($0 < \beta \leq 1$)
  Probanden, die diesem Niveau zugeordnet werden, verfügen über Grundkenntnisse der Gewichtskraft. Sie kennen die Coulombsche Theorie der Reibung (Gleitreibung) und Momente in nichtzentralen Kräftegruppen. Darüber hinaus können sie Aufgaben zu Fachwerken und den darin enthaltenen Stabkräften lösen.

- Niveau II ($1 < \beta \leq 2$)
  Weitere Grundlagen der Technischen Mechanik wie die Newtonschen Axiome sind bekannt. Die Probanden sind fähig, Themen der mechanischen Arbeit (Beschleunigung, Geschwindigkeit, Prinzip der virtuellen Arbeit und statische Belastung einer linearen Feder) zu behandeln. Die Bestimmung des Schwerpunkts, die Gleichgewichtsbedingungen für nicht-zentrale Kräftegruppen und die Coulombsche Theorie der Reibung in Bezug auf Haftreibung sind ebenfalls bekannt.

- Niveau III ($2 < \beta \leq 3$)
  Auf diesem Niveau können die Probanden den Begriff der Kraft definieren, die SI-Einheit der Arbeit reproduzieren und mit dem Kraft-Zeit-Diagramm arbeiten.

- Niveau IV ($3 < \beta \leq 4$)
  Probanden des vierten Niveaus kennen das Beschleunigungs-Zeit-Diagramm und können die Grundlagen der Technischen Mechanik

(SI-Einheit der Kraft, Newtonsche Axiome und Energieformen) reproduzieren.

- Niveau V ($\beta > 4$)
  Auf dem höchsten Niveau sind die Probanden in der Lage, Items zu Lagerreaktionen (Auflagerreaktionen) und zur Normalkraft in zentralen Kräftesystemen zu lösen.

Abgesehen von diesen Inhalten können keine schwierigkeitsbestimmende Merkmale der einzelnen Niveaus identifiziert werden. Es ist unklar, welche weiteren Merkmale einen Einfluss auf die Itemschwierigkeit haben. Das Studiendesign von FUNDAMENT und ALSTER ist als Large-Scale-Assessments nicht in der Lage, entsprechende Merkmale aufzudecken. Weiterführende Untersuchungen mit anderen wissenschaftlichen Beobachtungsmethoden wie Interviews oder ›lautes Denken‹ (Bortz & Schuster, 2010) können hier hilfreich sein.

Ein Vergleich der inhaltlichen Beschreibung der Niveaus mit dem Kernlehrplan Physik für die Sekundarstufe II für Gymnasien und Gesamtschulen in NRW (MSB NRW, 2022), kann die Herkunft des Wissens zumindest teilweise erklären. So finden sich im Kernlehrplan im Inhaltsfeld Grundlagen der Mechanik (Grundkurs) folgende inhaltliche Schwerpunkte:

- Kinematik
  - gleichförmige und gleichmäßig beschleunigte Bewegung
  - freier Fall
  - waagerechter Wurf
  - vektorielle Größen
- Dynamik
  - Newton‘sche Gesetze
  - beschleunigende Kräfte
  - Kräftegleichgewicht
  - Reibungskräfte

- Erhaltungssätze
  - Impuls
  - Energie (Lage-, Bewegungs- und Spannenergie)
  - Energiebilanzen
  - Stoßvorgänge (MSB NRW, 2022).

Die Inhalte des Niveau 0 sind im Kernlehrplan enthalten, ebenso die Inhalte des Niveaus I, wobei die Fachwerke und die zugehörigen Stabkräfte nicht Bestandteil des Kernlehrplans sind. Die im Niveau 2 angesiedelten Inhalte der mechanischen Arbeit (Prinzip der virtuellen Arbeit und statische Belastung einer linearen Feder) sowie die Bestimmung des Schwerpunkts und die Gleichgewichtsbedingungen für nicht-zentrale Kräftegruppen sind im Kernlehrplan nicht aufgeführt. Die grundlegenden Definitionen aus Niveau 3 sind ebenfalls (indirekt) im Kernlehrplan enthalten, jedoch nicht das Kraft-Zeit-Diagramm. Gleiches gilt für das Beschleunigungs-Zeit-Diagramm des vierten Niveaus. Die Lagerreaktionen, die in Niveau 5 erwähnt werden, sind nicht Teil des Kernlehrplans. Es zeigt sich, dass einige Inhalte durch das Kerncurriculum abgedeckt werden und somit das Wissen aus der Schule stammen kann. Diese Inhalte werden jedoch erst im Physikunterricht in der Sekundarstufe II vermittelt. Daher wird mit dem Kruskal-Wallis-Test, der Gruppenunterschiede für Rangdaten ermittelt (Eid et al., 2017), überprüft, ob die Wahl des Physik-Kurses in der Sekundarstufe II einen Einfluss auf die Kompetenzniveaus am zweiten MZP hat.

Der Kruskal-Wallis-Test ist die nicht-parametrische Alternative zur einfaktoriellen Varianzanalyse (ANOVA) (Field et al., 2012). Er setzt eine stetige Verteilung des Merkmals in der Grundgesamtheit, Stichprobenunabhängigkeit und eine homogene Streuung des Merkmals in den Subpopulationen voraus (Eid et al., 2017). Diese Bedingungen sind für die vorliegenden Daten erfüllt. Die Berechnung erfolgt in ›R‹ (R Core Team, 2023) mit dem ›R‹-Paket ›stats‹ (R Core Team, 2023) und der Funktion ›kruskal.test‹. Der Test bestätigt, dass es signifikante Unterschiede ($H(2) = 43.24$, $p < .001$) zwischen der Wahl des Physik-Kurses in der Sekundarstufe II und dem erreichten Niveau an MZP 2 gibt. Jedoch sagt dieses Ergebnis noch nichts darüber aus,

zwischen welchen Gruppen die Unterschiede auftreten. Daher wird eine nicht-parametrische Post-hoc-Berechnung mit Hilfe des gepaarten Wilcoxon-Rangsummentest durchgeführt (Field et al., 2012). Diese Art des Post-hoc-Tests führt zu einer α-Fehler-Kumulierung, die durch eine α-Adjustierung korrigiert werden kann (Eid et al., 2017; Field et al., 2012). In dieser Arbeit wird die konservative Bonferroni-Adjustierung verwendet (siehe Eid et al., 2017; Field et al., 2012). Diese Berechnung erfolgt wiederum mit dem ›R‹-Paket ›stats‹ (R Core Team, 2023), die Funktion lautet ›pairwise.wilcox.test‹. Die zugehörige Effektstärke r (siehe Kap. 6.2.1) wird mit dem ›R‹-Paket ›rstatix‹ (Kassambara, 2023) und der Funktion ›wilcox_effsize‹ ermittelt. Die berechneten Effektstärken sind in Tab. 6.31 aufgelistet.

Tabelle 6.31
Effektstärke des gepaarten Wilcoxon-Rangsummentests - Unterschiede zwischen der Wahl des Physik-Kurses in der Sekundarstufe II und dem erreichten Niveau an MZP 2 im Testinstrument der Technische Mechanik

	LK	GK
GK	0.23^{**}	
nicht belegt	0.35^{***}	0.21^{***}

Anmerkungen. MZP = Messzeitpunkt; LK = Leistungskurs; GK = Grundkurs.
$^{**}p < .01$. $^{***}p < .001$.

Es zeigt sich also, dass die Kurswahl im Fach Physik in der Oberstufe einen signifikanten Einfluss auf das Kompetenzniveau zu Studienbeginn im Testinstrument der Technischen Mechanik hat. Während zwischen der Nichtbelegung und der Wahl eines Grundkurses ein geringer Effekt besteht, zeigt sich dieser auch zwischen der Wahl eines Grundkurses und eines Leistungskurses. Ein mittlerer Effekt zeigt sich zwischen der Nichtbelegung und der Wahl eines Leistungskurses Physik. Dementsprechend kann ein höheres Kompetenzniveau zu Beginn des ersten Fachsemesters erreicht werden, wenn in der Sekundarstufe II ein Grundkurs oder noch besser ein Leistungskurs Physik belegt wurde.

Zu prüfen ist auch, ob die Abschlussnoten (in Punkten) im Fach Physik in der Oberstufe einen Einfluss auf das Kompetenzniveau ha-

ben. Es zeigt sich jedoch kein signifikanter Unterschied zwischen den Niveaus ($F(4, 90) = 0.60$, $p = .665$). Es scheint also ausreichend zu sein, einen Physik-Kurs in der Sekundarstufe II zu belegen, um ein höheres Kompetenzniveau zu erreichen, da die erreichte Note (am Ende der Sekundarstufe II) keinen Einfluss zeigt.

Darüber hinaus wird der Einfluss weiterer demographischer Variablen (Alter, Muttersprache, Bildungsherkunft, Art der Hochschulzugangsberechtigung und Note der HZB) auf das erreichte Kompetenzniveau an MZP 2 untersucht. Bis auf das Alter der Probanden haben alle Variablen einen Einfluss auf das Kompetenzniveau. Da sich nur ein Proband in Niveau V befindet, wurde dieses Niveau nicht in die Berechnungen einbezogen. Signifikante Unterschiede zwischen den Gruppen zeigen sich hinsichtlich der Art des Erwerbs der Hochschulzugangsberechtigung ($H(4) = 12.72$, $p = .013$). Der Unterschied zeigt sich mit einem kleinen Effekt zwischen Niveau 0 und Niveau IV ($p < .001$, $|r = |.18$), also zwischen dem niedrigsten und dem höchsten Niveau. Dementsprechend finden sich in dem höchsten Niveau mehr Probanden, die ihre Hochschulzugangsberechtigung an einem Gymnasium erworben haben, als in dem niedrigsten Niveau. Tatsächlich haben alle Probanden des Niveaus IV ihre Hochschulzugangsberechtigung durch das Abitur an einem Gymnasium erworben, während es in Niveau 0 nur 52% sind (38% an einer Gesamtschule). Auch bei der HZB-Note gibt es signifikante Unterschiede zwischen den Niveaus ($F(4, 467) = 16.10$, $p < .001$). Zwischen Niveau 0 und den Niveaus II, III und IV bestehen signifikante Unterschiede mit hoher Effektstärke ($p < .001$, $0.93 < d < 1.47$). Auch zwischen Niveau I und den Niveaus II ($p < .001$, $d = 0.77$) und IV ($p < .001$, $d = 1.23$) liegen signifikante Unterschiede mit mittlerer bzw. hoher Effektstärke vor. Daraus lässt sich ableiten, dass eine bessere HZB-Note zu einem höheren Kompetenzniveau führt.

Auch der Migrationshintergrund und die Bildungsherkunft haben einen Einfluss auf das Kompetenzniveau an MZP 2. So lassen sich signifikante Unterschiede zwischen den Niveaus feststellen, die darauf zurückzuführen sind, ob die Probanden Deutsch als Muttersprache haben oder nicht ($H(4) = 51.32$, $p < .001$). Das Niveau 0 unterscheidet sich signifikant mit einem kleinen oder mittleren Effekt von den Niveaus

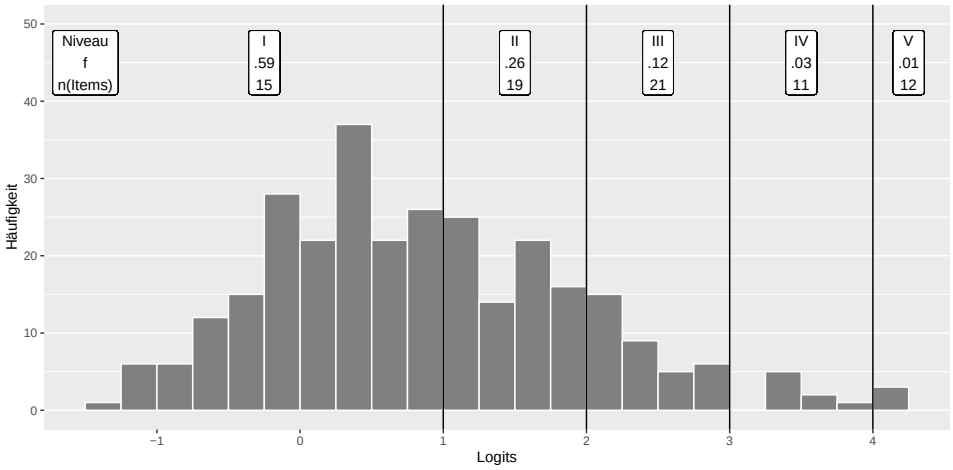
I, II und III ($p < .001$, $0.21 < r < 0.32$). Daraus lässt sich schließen, dass Probanden mit einer anderen Muttersprache als Deutsch eher die niedrigeren Niveaus (0 und I) erreichen. Bezüglich der Bildungsherkunft zeigen sich ebenfalls signifikante Unterschiede in den Kompetenzniveaus ($H(4) = 50.06$, $p < .001$). Mit einer geringen bis mittleren Effektstärke unterscheidet sich das Niveau 0 signifikant von den Niveaus I, II, III und IV ($p < .001$, $0.22 < r < 0.31$). Darüber hinaus gibt es signifikante Unterschiede mit einem mittleren Effekt zwischen Niveau I und den beiden höchsten Niveaus (III und IV) ($p < .001$, $|r| = 0.30$ bzw. 0.31). Somit lässt ein niedriger Bildungshintergrund auch ein niedriges Kompetenzniveau in der Technischen Mechanik zu Beginn des Studiums erwarten.

MZP 3 Analog zu MZP 2 wird an MZP 3 das Niveaumodell des Testinstruments der Technischen Mechanik entwickelt. Bei diesem MZP konnte jedoch eine Grenze weniger identifiziert werden. Es gibt nur noch vier Grenzen (bei den Punkten 1, 2, 3 und 4 auf der Logit-Skala), die wiederum zu fünf Kompetenzniveaus führen, die als I, II, III, IV und V bezeichnet werden. Das Niveau 0, welches es noch an MZP 2 gab, gibt es bei dem dritten MZP nicht mehr, da es keine Items mit einer entsprechenden Itemschwierigkeit von $\beta \leq 0$ gibt. Sieben Items erreichen Itemschwierigkeiten (bei $P(u_{ij} = 1) = 80\%$) von $\beta > 6$ und werden daher im Niveaumodell nicht berücksichtigt. Somit verbleiben 78 Items im Modell, welches in Abb. 6.15 dargestellt ist. Die Skalierung der y-Achse wurde beibehalten, um einen visuellen Vergleich zu ermöglichen. Die Säulenbreite wurde ebenfalls auf 0.25 Logits gesetzt. Lediglich die x-Achse wurde auf der linken Seite etwas verkürzt, da kein Proband eine Personenfähigkeit von $\theta < -1.5$ aufweist. Die Verteilung ist erneut rechtsschief.

Es ist zu erkennen, dass die Niveaus I, II und III (prozentual) deutlich an Probanden gewonnen haben. Niveau I steigt von 29% auf 59%, dies ist allerdings auch damit verbunden, dass Niveau 0 nicht mehr existiert. Aber auch die Zuwächse der Niveaus II und III um 16% bzw. 9% sind deutlich erkennbar.

Interessant ist an dieser Stelle, wie diese Zuwächse zustande kommen bzw. welche Niveaus die Probanden an MZP 2 hatten und welches

Abbildung 6.15
Niveaumodell des Testinstruments der Technischen Mechanik an MZP 3



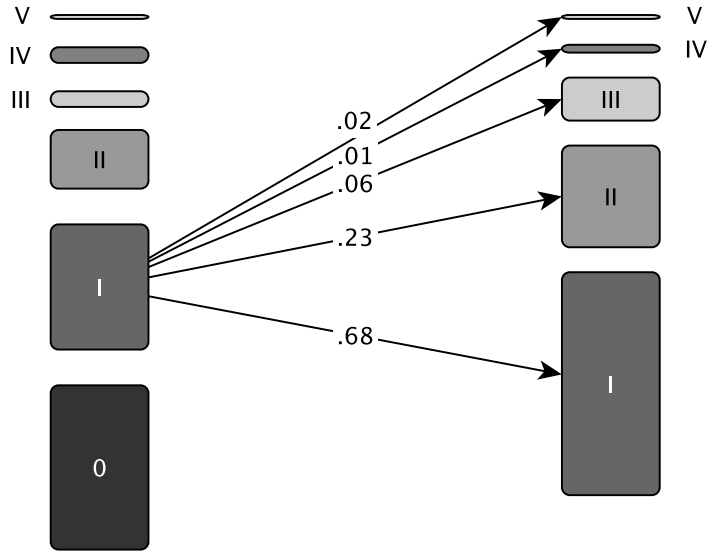
Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.
Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

Niveau sie am dritten MZP erreichen konnten. Dazu wird eine ›Niveauwanderung‹ untersucht. Diese ist für das Niveau I an MZP 2 in Abb. 6.16 visualisiert.

Diese Abbildung ist so zu verstehen, dass die Pfeile anzeigen, welches Niveau ein Proband an MZP 3 erreicht, wenn er zuvor das erste Niveau an MZP 2 erreicht hat. Die Prozentzahlen an den Pfeilen geben die relative Häufigkeit an, z.B. sind 68% der Probanden in Niveau I geblieben, während 26% den ›Aufstieg‹ in Niveau II geschafft haben. Insgesamt haben 32% den Aufstieg in ein höheres Niveau gemeistert. Hierbei ist zu beachten, dass nicht alle Probanden an beiden MZP teilgenommen haben und somit die Stichprobengröße bei dieser ›Niveauwanderung‹ kleiner ist als bei der Einzelbetrachtung der MZP. Zur besseren Übersicht ist in Abb. 6.16 nur Niveau I dargestellt, die weiteren ›Niveauwanderungen‹ sind in Tab. 6.32 zu finden.

In der ersten Spalte der Tab. 6.32 sind die Ausgangsniveaus an MZP 2 angegeben. Jede Zeile gibt dementsprechend die relative Häufigkeit an, mit der sich die Probanden eines Niveaus an MZP 2 auf die Niveaus am folgenden MZP 3 verteilen. Die hervorgehobenen Werte

Abbildung 6.16
›Niveauwanderung‹ des Testinstruments der Technischen Mechanik von MZP 2 (Niveau I) nach MZP 3



Anmerkungen. MZP = Messzeitpunkt; angegeben sind die relativen Häufigkeiten.

Tabelle 6.32
›Niveauwanderung‹ von MZP 2 nach MZP 3 - Testinstrument Technische Mechanik

MZP 2	MZP 3				
Niveau	I	II	III	IV	V
0	.71	.23	.06		
I	**.68**	.23	.06	.01	.02
II	.24	**.44**	.22	.07	.02
III		.36	**.45**	.09	.09
IV	.36	.18	.36	**.09**	
V				1.00	

Anmerkungen. MZP = Messzeitpunkt.
Die Abweichungen vom Wert 100 sind auf Rundungsfehler zurückzuführen.

geben den Prozentsatz der Probanden an, die in ihrem jeweiligen Niveau verbleiben. Werte überhalb der Diagonalen stellen die Probanden dar, die ein höheres Niveau erreichen konnten. Werte unterhalb der Diagonalen stehen dagegen für eine Verschlechterung und damit für ein niedrigeres Kompetenzniveau. Die Zeilensummen ergeben immer 100%, Abweichungen sind auf Rundungsfehler zurückzuführen. Knapp 70% der Probanden sind in Niveau I verblieben, in Niveau II und III sind es jeweils nur rund 45%. 71% sind von Niveau 0 in Niveau 1 aufgestiegen, wobei zu beachten ist, dass es kein Niveau 0 mehr gibt und somit Niveau 1 die niedrigste Einstufung darstellt. Jeweils ca. 23% der Probanden aus den Niveaus II und III konnten eine Stufe höher aufsteigen, während aus Niveau III nur 9% den Aufstieg in Niveau IV schafften. Einigen Probanden gelang es sogar, ein oder mehrere Niveaus zu überspringen. Diese positiven Beobachtungen deuten darauf hin, dass die Probanden ihr Fachwissen in der Technischen Mechanik während des ersten Fachsemesters verbessern konnten. Es gibt aber auch gegenteilige Beobachtungen. So haben sich in den Niveaus II, III und IV zwischen 24% und 36% der Probanden um ein Niveau verschlechtert – in Niveau V sind es sogar 100%, da der einzige Proband an MZP 3 nun das vierte Niveau besetzt. Der Abstieg um ein Niveau ist auch inhaltlich durchaus plausibel, da es auch zu Schwankungen der Testleistung am Testtag kommen kann. Allerdings stellen die Abstiege von Niveau IV in Niveau II und sogar I mit insgesamt 54% in dieser Größenordnung eine Besonderheit dar. Eine Erklärung hierfür könnte ein Motivationsverlust der Probanden bei der Bearbeitung der Testinstrumente sein. Da in FUNDAMENT zwischen 59 und 70 Items und in ALSTER sogar 118 bzw. 148 Items in den Testinstrumenten der Technischen Mechanik eingesetzt wurden (siehe Tab. 5.2), erscheint ein Motivationsverlust bei der Bearbeitung der Items durchaus möglich.

Analog zum vorangegangenen MZP erfolgt die inhaltliche Beschreibung der Niveaus anhand der zugeordneten Items:

- Niveau I ($\beta \leq 1$)
  Die Einstufung in das niedrigste Niveau bescheinigt den Probanden Kenntnisse über Kraftsysteme. Items zu zentralen (Resultierende und Zerlegung von Kräften) und nicht-zentralen Kraftsystemen (Resultierende sowie Momente und Gleichgewichtsbedingungen) können gelöst werden. Auch Fachwerke, die zu mehrteiligen Tragwerken gehören, sind in Bezug auf Nullstäbe, Stabkräfte oder Gelenke bekannt. Ebenso können Lager- und Auflagerreaktionen besimmt werden.

- Niveau II ($1 < \beta \leq 2$)
  Probanden des zweiten Niveaus sind in der Lage, die Eigenschaften von Kräften und die Newtonschen Axiome zu reproduzieren. Darüber hinaus sind sie in der Lage, Schwerpunktsberechnungen durchzuführen und Lagerreaktion in Bezug auf Auflagerreaktionen und Lagertypen ebener Tragwerke zu untersuchen. Items zu Schnittgrößen (am geraden Balken und von Rahmensystemen) sowie zur Coloumbschen Theorie der Reibung (Haftreibung) können gelöst werden.

- Niveau III ($2 < \beta \leq 3$)
  Auf Niveau III sind die Coloumbsche Theorie der Reibung (Gleitreibung) sowie Schnittgrößen bei Beanspruchung durch Einzelkräfte und Einzelmomente bekannt. Außerdem sind die Probanden imstande Items zur mechanischen Arbeit (Prinzip der virtuellen Arbeit und statische Belastung einer linearen Feder) zu lösen.

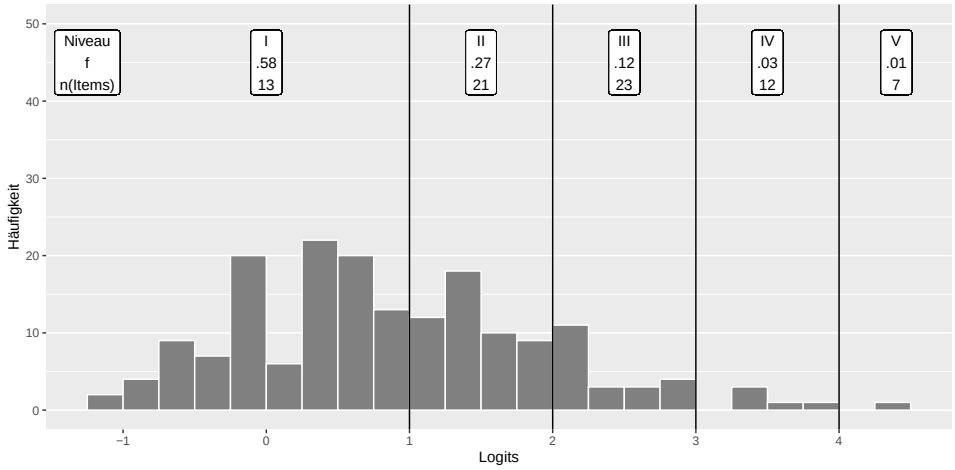
- Niveau IV ($3 < \beta \leq 4$)
  Die Probanden des vierten Niveaus können Items lösen, die im Vergleich zu den Items der vorhergehenden Niveaus einen höheren Komplexitätsgrad aufweisen.

- Niveau V ($\beta > 4$)
  Auf dem höchsten Niveau sind die Probanden in der Lage, grundlegende Energieformen und Kraft-Zeit-Diagramme zu reproduzieren. Weiterhin können sie die statische Bestimmtheit mehrteiliger Tragwerke ermitteln.

Im Vergleich zu MZP 2 ist festzustellen, dass die Inhalte der Technischen Mechanik anderen Niveaus zugeordnet sind als zuvor.

MZP 4 Die Entwicklung des Niveaumodells für das Testinstrument der Technischen Mechanik an MZP 4 folgt dem gleichen Prinzip wie die beiden vorangeganene MZP. Analog zu MZP 3 konnten insgesamt fünf Kompetenzniveaus identifiziert werden, die durch vier Grenzen (bei den Punkten 1, 2, 3 und 4 auf der Logit-Skala) begrenzt werden. Die Niveaus werden wiederum mit I, II, III, IV und V bezeichnet. Drei Items konnten im Niveaumodell nicht berücksichtigt werden, da ihre Itemschwierigkeiten (bei $P(u_{ij} = 1) = 80\%$) bei $\beta > 6$ liegen. Insgesamt befinden sich somit noch 76 Items an MZP 4 in dem Niveaumodell. Das zugehörige Diagramm (Abb. 6.17) folgt dem gleichen Aufbau wie die vorherigen Diagramme (gleiche Skalierung der y-Achse und Säulenbreite von 0.25 Logits).

Abbildung 6.17
Niveaumodell des Testinstruments der Technischen Mechanik an MZP 4



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.
Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

Die rechtsschiefe Verteilung zeigt, dass 59% der Probanden nur das niedrigste Niveau erreichen. 26% werden dem zweiten Niveau zugeordnet, während 16% die Niveaus II bis V erreichen. Im Vergleich zu MZP 3

(siehe Abb. 6.15) ergibt sich nahezu die gleiche Verteilung – lediglich die Niveaus I und II unterscheiden sich um jeweils .01.

Es ist daher durchaus interessant, in einer erneuten ›Niveauwanderung‹ (MZP 3 nach MZP 4) zu prüfen, ob die Probanden in ihren Niveaus weitestgehend verbleiben. Die entsprechenden Daten sind in Tab. 6.33 zusammengestellt.

Tabelle 6.33
›Niveauwanderung‹ von MZP 3 nach MZP 4 - Testinstrument Technische Mechanik

MZP 3	MZP 4				
Niveau	I	II	III	VI	V
I	**.94**	.06			
II	.29	**.61**	.11		
III	.18	.36	**.45**		
IV		.14	.43	**.29**	.14
V			.67	.33	

Anmerkungen. MZP = Messzeitpunkt.
Die Abweichungen vom Wert 100 sind auf Rundungsfehler zurückzuführen.

Die Ausgangsniveaus an MZP 3 sind in der ersten Spalte der Tab. 6.33 angegeben. In den weiteren Spalten sind die erreichten Niveaus mit den entsprechenden relativen Häufigkeiten angegeben. 94% der Probanden verbleiben in dem niedrigsten Niveau, nur 6% konnten sich auf Niveau II verbesern. Weitere 11% bzw. 14% schafften auch den Aufstieg um eine Stufe von Niveau II in III bzw. Niveau IV in V. Damit sind die Aufstiege deutlich seltener und begrenzter (höchsten um eine Stufe), als von MZP 2 nach MZP 3. Eine Verschlechterung des Niveaus ist in allen Niveaus – mit Ausnahme des niedrigsten Niveaus – zu beobachten. Während sich in Niveau II knapp ein Drittel der Probanden um eine Stufe verschlechtert, sind es in den Niveaus III und IV mehr als die Hälfte, 18% bzw. 14% sogar um zwei Stufen. Vom höchsten Niveau V verschlechtern sich alle Probanden.

Die den Niveaus zugeordneten Items werden erneut zur inhaltlichen Beschreibung der Niveaus herangezogen. Da nun auch Items aus der Veranstaltung Technischen Mechanik 2 verwendet werden, orientiert

sich die Benennung innerhalb der Niveaus an den Skripten der beiden Vorlesungen von Schröder (2020, 2021).

- Niveau I ($\beta \leq 1$)
  Probanden des niedrigsten Niveaus sind in der Lage, die Gewichtskraft zu beschreiben. Sie kennen mehrteilige Tragwerke (Fachwerke und Nullstäbe) und können die Resultierenden von zentralen Kraftsystemen sowie Lagerreaktionen (Auflagerreaktionen) bestimmen. Sie können auch Items zur nicht-zentralen Kräftegruppe (Momente, Resultierende und Gleichgewichtsbedingungen), Schnittgrößen von Rahmensystemen und zur Torsion lösen.

- Niveau II ($1 < \beta \leq 2$)
  Auf der zweiten Ebene sind die Probanden fähig, den Grundbegriff der Masse und die Newtonschen Axiomen zur reproduzieren. Sie können die Stabkräfte eines mehrteiligen Tragwerks und die mechanische Arbeit (Geschwindigkeit und statische Belastung einer linearen Feder) bestimmen sowie die Coulombsche Theorie der Reibung (Haftreibung und Gleitreibung) anwenden. In einem zentralen Kräftesystem können Items zur Zerlegung von Kräften gelöst werden und in der Statik deformierbarer Körper können Spannungs- und Dehnungszustände sowie die entsprechenden Schnittgrößen bestimmt werden.

- Niveau III ($2 < \beta \leq 3$)
  Auf Niveau III sind das Kraft-Zeit-Diagramm, die Schnittgrößen am geraden Balken und die Beschleunigung der mechanischen Arbeit bekannt. Die Randbedingungen von Lagerreaktionen und der Schwerpunkt eines Körpers können ebenfalls bestimmt werden.

- Niveau IV ($3 < \beta \leq 4$)
  Probanden, die Niveau IV erreicht haben, sind mit dem Konzept der Flächenträgheitsmomente, der technischen Biegetheorie und der Querdehnung in der Statik deformierbarer Körper vertraut.

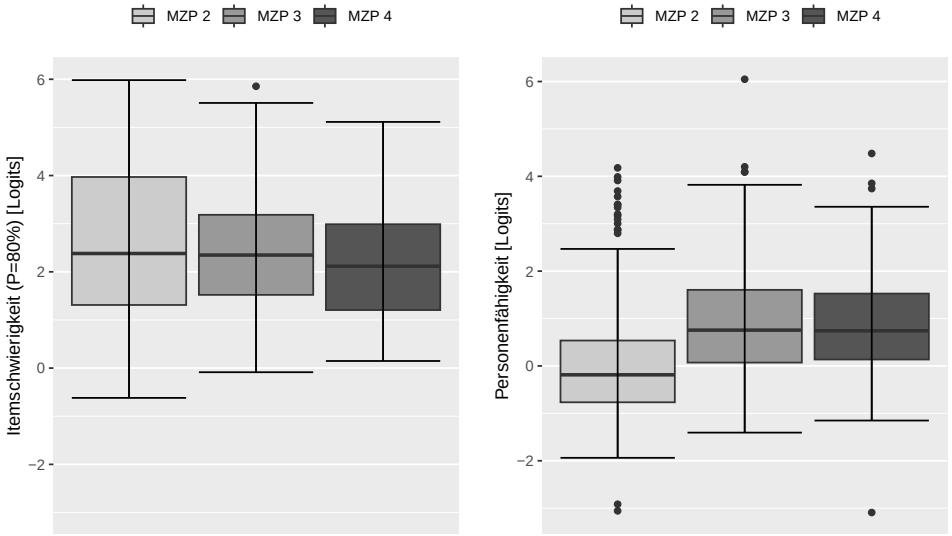
- Niveau V ($\beta > 4$)
  Auf dem höchsten Niveau sind die Probanden in der Lage, das

Prinzip der virtuellen Kräfte anzuwenden und die statische Bestimmtheit von mehrteiligen Tragwerken zu ermitteln.

Damit ist die Niveaumodellierung für die Testinstrumente der Technischen Mechanik für alle drei MZP abgeschlossen. Abschließend können die Ergebnisse miteinander verglichen werden.

Vergleich der MZP 2 bis 4 Zum Vergleich der drei MZP der Technischen Mechanik werden zunächst die Itemschwierigkeiten ($P(u_{ij} = 1) = 80\%$) und die erreichten Personenfähigkeiten gegenübergestellt (Abb. 6.18). Die abgebildeten Diagramme zeigen Box-Whisker-Diagramme für die einzelnen MZP, auf der Y-Achse sind die Logits aufgetragen.

Abbildung 6.18
Box-Whisker-Diagramm der Itemschwierigkeit ($P(u_{ij} = 1) = 80\%$) und der Personenfähigkeit der Testinstrumente der Technischen Mechanik an den MZP 2 bis 4



Anmerkungen. MZP = Messzeitpunkt.

Zur Erinnerung: Die Größe des Kastens ist durch das erste und dritte Quartil festgelegt und die waagerechte Linie in dem Kasten stellt den Median dar (siehe Kap. 6.1.1). Der Median bleibt konstant über einem Logit-Wert von 2, mit einer minimalen Verschiebung nach unten am letzten MZP. Die Größe des Kastens nimmt von MZP 2 nach MZP 3

stark ab. Um statistisch zu überprüfen, ob sich die drei MZP signifikant unterscheiden, wird ein parametrischer Test mit der einfaktoriellen ANOVA berechnet. Während der zuvor verwendete Kruskal-Wallis-Test mindestens ordinalskalierte Daten voraussetzt, müssen die Daten für die ANOVA mindestens intervallskaliert sein (Eid et al., 2017). Weitere Voraussetzungen für die ANOVA sind die Unabhängigkeit der Gruppen sowie Normalverteilung und Varianzhomogenität (Eid et al., 2017; Field et al., 2012). Nach Field et al. (2012) ist die ANOVA aber auch robust gegenüber Verletzungen der Normalverteilung. Die ANOVA analysiert die Mittelwertunterschiede zwischen den Gruppen (unabhängige Variable). Bei einem signifikanten Ergebnis wird die Nullhypothese, dass die Varianzen zwischen den Gruppen gleich sind, verworfen. Die Berechnung erfolgt in ›R‹ (R Core Team, 2023) mit dem ›R‹-Paket ›stats‹ (R Core Team, 2023) und der Funktion ›aov‹. Die Effektstärke der ANOVA ist weniger von Interesse, vielmehr wird bei signifikanten Ergebnissen ein Post-hoc-Test durchgeführt, wie bereits beim Kruskal-Wallis-Test beschrieben. Bei der ANOVA ist dies der gepaarte t-Test, auch hier wird eine α-Adjustierung vorgenommen. Mit dem ›R‹-Paket ›stats‹ (R Core Team, 2023) und der Funktion ›pairwise.t.test‹ wird der gepaarte t-Test berechnet und anschließend mit der Funktion ›cohens_d‹ aus dem ›R‹-Paket ›effectsize‹ (Ben-Shachar et al., 2020) die Effektstärke Cohen's d. Die Werte zur Interpretation von d finden sich in Kap. 5.5. Es ist zu beachten, das die Funktion ›cohens_d‹ nur die Beträge von Cohen's d ausgibt, sodass das genaue Vorzeichen unklar ist und nur inhaltlich interpretiert werden kann. Die Voraussetzungen der ANOVA sind erfüllt und die Berechnung zeigt keine signifikanten Unterschiede in den Itemschwierigkeiten ($F(2, 216 = 1.09$, $p = .339$). Da kein signifikantes Ergebnis vorliegt, kann auf die Berechnung des Post-hoc-Tests verzichtet werden.

Der direkte Vergleich der Itemschwierigkeiten und der Personenfähigkeiten zeigt, dass die Itemschwierigkeiten auf der y-Achse in positive Richtung verschoben sind. Der Grund hierfür liegt in der gewählten Lösungswahrscheinlichkeit von $P(u_{ij} = 1) = 80\%$, die zu entsprechend höheren Itemschwierigkeiten führt und diese Verschiebung bedingt. Für $P(u_{ij} = 1) = 50\%$ liegen die mittleren Itemschwierigkeiten zwischen

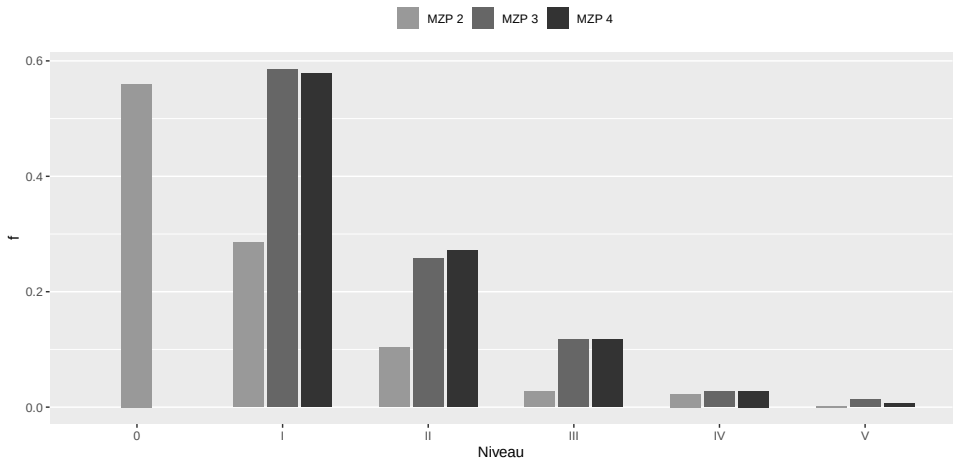
0.21 und 0.92 (siehe Kap. 6.2.3.1).

Abb. 6.18 zeigt, dass der Median der Personenfähigkeit an MZP 2 am kleinsten ist (unter 0 Logits). An den MZP 3 und 4 gibt es keine Unterschiede im Median. Die ANOVA zeigt signifikante Unterschiede zwischen den MZP ($F(2, 960) = 78.15$, $p < .001$). Die Post-hoc-Berechnung bestätigt die aus der Abbildung abgeleiteten Aussagen. Es gibt keine signifikanten Unterschiede zwischen MZP 3 und 4, jedoch zwischen MZP 2 und MZP 3 ($p < .001$, $d = -0.80$) sowie MZP 4 ($p < .001$, $d = -0.81$) mit jeweils hoher Effektstärke.

Somit kann zusammengefasst werden, dass sich die Itemschwierigkeiten zwischen den drei MZP nicht unterscheiden, die Personenfähigkeiten an MZP 2 jedoch signifikant niedriger sind als an den beiden anderen MZP. Weiterhin zeigt sich, dass es keine signifikante Steigerung der Personenfähigkeiten am dritten und vierten MZP gibt, also keine Verbesserung im zweiten Fachsemester.

Diese Aussagen können auch durch die Betrachtung der Kompetenzniveaus bestätigt werden. Dazu ist in Abb. 6.19 ein Säulendiagramm dargestellt, in dem die Höhe der Säulen den jeweiligen relativen Häufigkeiten der Niveaus an allen drei MZP entspricht.

Abbildung 6.19
Verteilung der Kompetenzniveaus der Testinstrumente der Technischen Mechanik an den MZP 2 bis 4



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit.

Im direkten Vergleich aller drei MZP wird deutlich, dass jeweils das niedrigste Niveau (MZP 2: Niveau 0, MZP 3 und 4: Niveau I) die meisten Probanden aufweist. Außerdem ist ein deutlicher Anstieg von MZP 2 nach MZP 3 über alle Niveaus zu erkennen. Dieser Anstieg lässt sich auch inhaltlich begründen (siehe Kap. 6.3.1). Es gibt jedoch keine Veränderung zwischen MZP 3 und MZP 4. Der Kruskal-Wallis-Test weist auf signifikante Unterschiede zwischen den Niveaus hin ($H(2) = 281.58$, $p < .001$). Die Post-hoc-Berechnung zeigt keine signifikanten Unterschiede zwischen MZP 3 und MZP 4. Allerdings unterscheiden sich MZP 2 und 3 wiederum mit einem hohen ($p < .001$, $|r| = 0.52$) sowie MZP 2 und 4 mit einem mittleren Effekt ($p < .001$, $|r| = 0.48$). Als Fazit kann festgehalten werden, dass sich die Itemschwierigkeiten über die MZP nicht signifikant unterscheiden, wobei die Personenfähigkeiten von MZP 2 zu MZP 3 signifikant ansteigen, während von MZP 3 zu MZP 4 keine signifikanten Veränderungen festzustellen sind. Dies zeigt sich auch bei der Betrachtung der Kompetenzniveaus. Während sich an MZP 2 85% der Probanden in den Niveaus 0 und I befinden, sind an den beiden anderen MZP die Niveaus II und III stärker ausgeprägter.

Die Gründe für den Anstieg von MZP 2 zu MZP 3 wurden bereits in diesem Kapitel erläutert. Unklar ist allerdings, wie die nahezu konstanten Ergebnisse von MZP 3 und MZP 4 zu interpretieren sind. Da es keine signifikanten Unterschiede bei den Itemschwierigkeiten gibt, muss die Ursache bei den Probanden liegen. Eine Interpretationsmöglichkeit der konstanten Personenfähigkeiten wäre, dass die Probanden im zweiten Fachsemester keine Kenntnisse erworben haben. Dies ist jedoch stark zu bezweifeln, da im vierten MZP auch Items verwendet wurden, die speziell auf die Inhalte der Veranstaltung Technische Mechanik 2 im zweiten Fachsemester zugeschnitten sind. Außerdem bestehen die Niveaus an den beiden MZP nicht aus den gleichen Items. D.h. einige Items, die an MZP 3 eine höhere Itemschwierigkeit hatten, sind an MZP 4 leichter, aber auch der umgekehrte Fall ist möglich. Es kann also davon ausgegangen werden, dass die Probanden genügend Wissen zur Technischen Mechanik 2 erworben haben, um ihr bisherige Kompetenzniveau zu halten. Der Wissenserwerb reicht jedoch nicht aus, um in das nächsthöhere Niveau aufzusteigen. Weiterhin lösen einige Probanden

Items an MZP 4 nicht mehr, obwohl sie dieses Item am vorherigen MZP richtig gelöst haben. Dies deutet auf ›Vergessenseffekte‹ oder Motivationsproblemen bei der Bearbeitung der Testinstrumente hin.

Zusammenfassend lassen sich Prädiktoren für das erreichte Kompetenzniveau an MZP 2 herausheben. So erweist sich die Wahl des Physik-Kurses (keine Belegung, Grundkurs oder Leistungskurs) als signifikant. Probanden, die einen Leistungskurs belegt haben, erreichen ein höheres Niveau, als Probanden, die einen Grundkurs belegt haben, sind aber dennoch den Probanden überlegen, die keinen Physik-Unterricht in der Oberstufe hatten. Auch die Art des Erwerbs der Hochschulzugangsberechtigung hat einen signifikanten Einfluss auf das Kompetenzniveau. Probanden, die ihre Hochschulreife an einem Gymnasium erworben haben, erreichen höhere Kompetenzniveaus. Die Note der Hochschulzugangsberechtigung hat ebenfalls einen signifikanten Einfluss auf die Kompetenzniveaus, wobei schlechte Noten auf eine Einstufung in niedrigere Kompetenzniveaus hindeuten. Auch der Migrationshintergrund der Probanden zeigt sich in Form der Muttersprache als Prädiktor. Es gibt signifikante Unterschiede, ob die Probanden Deutsch oder eine andere Sprache als Muttersprache haben. Sofern Deutsch nicht die Muttersprache ist, wird in der Regel eines der niedrigeren Kompetenzniveaus erreicht. Schließlich führt auch die Betrachtung des Bildungshintergrunds zu signifikanten Unterschieden zwischen den Kompetenzniveaus. Ein niedriger Bildungshintergrund spricht für eine Zuordnung zu den niedrigen Niveaus, während ein hoher Bildungshintergründe auf ein hohes Kompetenzniveau hinweist.

Diese Prädiktoren sind für den Studienverlauf entscheidend, da es den Probanden in der Regel nicht gelingt, im Verlauf des Studiums ein höheres Kompetenzniveau in den Testinstrumenten der Technischen Mechanik zu erreichen und sie somit als Risikogruppe zu bezeichnen sind.

6.3.2 Rechenfähigkeit (FUNDAMENT)

Bereits in Kap. 6.2.1 wurde festgestellt, dass eine gemeinsame Skalierung der Testinstrumente zur Erfassung der Rechenfähigkeit für FUNDAMENT und ALSTER nicht zulässig ist und daher beide Studi-

en getrennt skaliert werden müssen. Dies hat zur Folge, dass auch die Niveaumodelle getrennt entwickelt werden müssen.

Zunächst erfolgt die Niveaumodellierung der Testinstrumente der Rechenfähigkeit für die Studie FUNDAMENT. Da diese vorliegenden Daten, wie bereits die Daten der Technischen Mechanik, bei der IRT-Skalierung die beste Passung für das 2pl-Birnbaum-Modell zeigen (siehe Kap. 6.2.3.2), muss vorab geprüft werden, ob die Rangfolge der Items gleich ist (siehe Kap. 6.3). Analog zur Niveaumodellierung der Technischen Mechanik wird die Rangfolge der Items durch den Pearson Korrelationskoeffizienten (r) beschrieben. Hierbei werden die Itemschwierigkeiten des 2pl-Birnbaum-Modells ($P(u_{ij} = 1) = 80\%$) mit den Itemschwierigkeiten des 1pl-Rasch-Modells ($P(u_{ij} = 1) = 50\%$) verglichen. Die Voraussetzungen für die Berechnung der Produkt-Moment-Korrelation nach Pearson (siehe Kap. 6.2.1) sind erfüllt. Die Ergebnisse für die MZP 2 bis 4 für die Testinstrumente der Rechenfähigkeit der Studie FUNDAMENT finden sich in Tab. 6.34.

Tabelle 6.34
Pearson Korrelationskoeffizienten der Itemschwierigkeiten der Testinstrumente der Rechenfähigkeit (FUNDAMENT) (2pl-Birnbaum-Modell [$P(u_{ij} = 1) = 80\%$] und 1pl-Rasch-Modell [$P(u_{ij} = 1) = 50\%$])

	t	*df*	*p*	*r*
MZP 2	10.81	18	<.001	.93
MZP 3	2.63	10	.025	.64
MZP 4	0.96	10	.357	.29

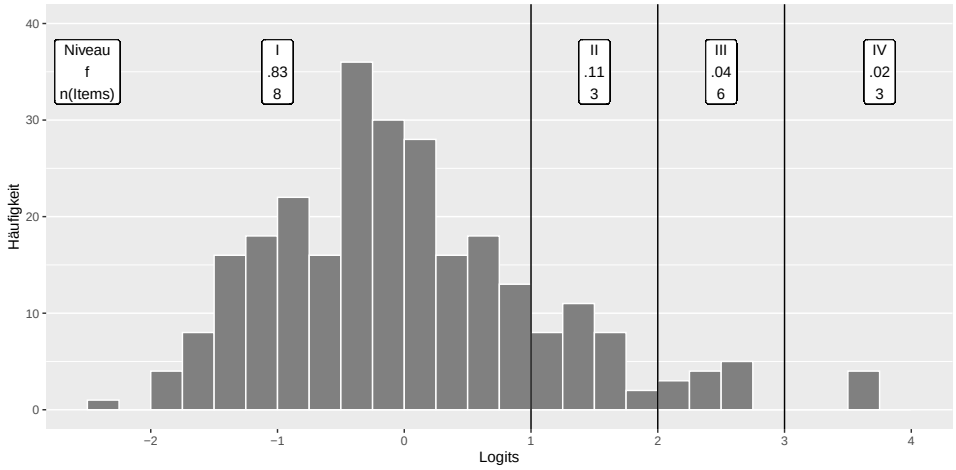
Anmerkungen. t = t-Wert; *df* = Freiheitsgrade; p = Wahrscheinlichkeit; r = Pearson Korrelationskoeffizient; MZP = Messzeitpunkt.

Hohe signifikante Korrelationen zeigen sich nur an den MZP 2 und 3, die somit für eine Niveaumodellierung verwendet werden können. An MZP 4 kann kein signifikanter Zusammenhang zwischen den Itemschwierigkeiten des 2pl-Birnbaum-Modells ($P(u_{ij} = 1) = 80\%$) und des 1pl-Rasch-Modells ($P(u_{ij} = 1) = 50\%$) festgestellt werden. Da Wilson (2003) die Überprüfung der Rangfolgen nicht als Kriterium für die Verwendung der Itemschwierigkeiten nennt, werden auch die Daten

des vierten MZP für die Niveaumodellierung verwendet.

MZP 2 Für MZP 2 lassen sich für das Testinstrument der Rechenfähigkeit in der Studie FUNDAMENT drei Niveaugrenzen identifizieren. Auf der Logit-Skala liegen die Grenzen bei den Punkten 1, 2 und 3. Aus diesen Grenzen lassen sich vier Kompetenzniveaus ableiten, die in Analogie zu den Testinstrumenten der Technischen Mechanik als I, II, III und IV bezeichnet werden. Niveau I umfasst Itemschwierigkeiten (bei $P(u_{ij} = 1) = 80\%$) von $\beta \leq 1$, Niveau II umfasst $1 < \beta \leq 2$ usw. Die maximale Itemschwierigkeit von $\beta < 6$ kommt nicht zum Tragen, da alle Items eine geringere Itemschwierigkeit aufweisen. Insgesamt gehen 20 Items in das Niveaumodell ein. Zur Darstellung der Niveaumodelle der Testinstrumente der Rechenfähigkeit werden wiederum Histogramme verwendet (Säulenbreite 0.25 Logits). Das resultierende Niveaumodell ist in Abb. 6.20 dargestellt.

Abbildung 6.20
Niveaumodell des Testinstruments der Rechenfähigkeit an MZP 2 - FUNDAMENT



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.

Auch in diesem Niveaumodell werden die Auswirkungen der festgesetzten hohen Lösungswahrscheinlichkeit deutlich. Während die höheren Kompetenzniveaus (II bis IV) jeweils einen Bereich von einem Logit

abdecken und entsprechend gut differenzieren können, umfasst das niedrigste Niveau einen Bereich von über 3 Logits. Da für das erste Niveau nur Items mit einer Itemschwierigkeit (bei $P(u_{ij} = 1) = 80\%$) von $0 < \beta \leq 1$ vorliegen, ist eine weitere Unterteilung nicht möglich, weil ansonsten ein weiteres Niveau ohne vorhandene Items inhaltlich nicht beschrieben werden könnte. Dennoch wären Items in diesem Bereich wünschenswert, um das unterste Niveau weiter abgrenzen zu können.

Die Verteilung der Probanden ist rechtsschief und die meisten Probanden (83%) werden dem niedrigsten Niveau zugeordnet. Auffällig ist, dass die höchsten Balken, d.h. die höchsten absoluten Zahlen, nicht an der Grenze zum Niveau II liegen, sondern im Bereich von -0.5 bis 0.25 Logits. Auch dies ist ein Hinweis darauf, dass eine weitere Niveaugrenze bei 0 sinnvoll wäre, da aber keine inhaltliche Interpretation möglich wäre, wird hiervon abgesehen. Das Niveau II stellt mit 11% der Probanden die zweitgrößte Gruppe dar, in den beiden höchsten Niveaus sind noch 6% der Probanden vertreten.

Während bei den Testinstrumenten der Technischen Mechanik inhaltlich nachvollziehbar ist, warum die Personenfähigkeiten am zweiten MZP niedriger sind (relativ betrachtet), kann die Argumentation für die Rechenfähigkeit nicht analog erfolgen. Die Items zur Rechenfähigkeit beschränken sich auf grundlegende mathematische Inhaltsbereiche (siehe Kap. 5.2) und sollten dementsprechend bereits in der Schule erworben worden sein. Somit können an dieser Stelle die in Kap. 2.2.1 postulierten Schwächen der Studienanfänger im mathematischen Bereich bestätigt werden. Es sei auch noch einmal darauf hingewiesen, dass eine mögliche positive Entwicklung in den folgenden MZP nicht auf die Mathematikveranstaltungen zurückzuführen ist, da diese grundlegenden Inhalte in den Veranstaltungen in der Regel nicht behandelt und vorausgesetzt werden.

Die Niveaus werden inhaltlich anhand der zugeordneten Items beschrieben. Auch die Inhalte der Rechenfähigkeit werden nur in einem Kompetenzniveau explizit genannt, allerdings können entsprechende Inhalte in höherer Komplexität in den höheren Niveaus erneut vorkommen.

- Niveau I ($\beta \leq 1$)
  Probanden in diesem Niveau beherrschen die Grundrechenarten und können Items lösen, bei denen es um Potenz- und Bruchrechnung (Addition, Multiplikation, Potenzen und Variablen) geht. Sie kennen den Satz des Pythagoras, können Stammfunktionen bilden und Gleichungen lösen.

- Niveau II ($1 < \beta \leq 2$)
  In diesem Niveau können Gleichungen höherer Komplexität gelöst werden.

- Niveau III ($2 < \beta \leq 3$)
  Auf diesem Niveau können die Probanden Ungleichungen lösen und Winkelfunktionen anwenden. In der Analysis können Ableitungen gebildet und Parabeln beschrieben werden.

- Niveau IV ($\beta > 3$)
  In dem höchsten Niveau sind die Probanden in der Lage, mit Wurzeln zu rechnen.

Alle in den Niveaus genannten Inhalte sind im Kernlehrplan Mathematik für die Sekundarstufe II für Gymnasien und Gesamtschulen in NRW (MSB NRW, 2023) zu finden. Da alle Schüler in der Sekundarstufe II einen Mathematikkurs belegen, kann nur untersucht werden, ob die Wahl eines Grund- oder Leistungskurses einen Einfluss auf das Niveau an MZP 2 hat. Hier zeigt sich ein signifikanter Unterschied mit einer geringen Effektstärke ($H(1) = 16.66$, $p < .001$, $|r| = 0.26$). Somit erreichen Probanden, die in der Schule einen Leistungskurs Mathematik besucht haben, ein höheres Kompetenzniveau als Probanden, die in der Sekundarstufe II einen Grundkurs belegt haben.

Im Folgenden wird untersucht, ob die Abschlussnoten (in Punkten) im Fach Mathematik in der Oberstufe einen Einfluss auf das Kompetenzniveau haben. Es zeigt sich ein signifikanter Unterschied zwischen den Niveaus ($F(3, 206) = 3.54$, $p = .016$). Dieser signifikante Unterschied besteht zwischen dem niedrigsten Niveau 0 und dem höchsten Niveau IV ($p < .001$, $d = -1.22$). Eine schlechte Note (am Ende der Sekundarstufe II) deutet also darauf hin, dass die Probanden auch in Bezug

auf ihre Rechenfähigkeit einem niedrigen Kompetenzniveau zugeordnet werden.

Bei der Kontrolle verschiedener demographischer Variablen (Alter, Muttersprache, Bildungsherkunft, Art der Hochschulzugangsberechtigung und Note der HZB) zeigen sich signifikante Unterschiede zwischen den Niveaus nur bei der Note der HZB ($F(3, 264) = 7.74$, $p < .001$). Das Niveau I unterscheidet sich signifikant mit einem mittleren Effekt von Niveau II ($p < .001$, $d = 0.51$) und mit einem großen Effekt von Niveau IV ($p < .001$, $d = 1.78$). Dies deutet darauf hin, dass schlechtere HZB-Noten zu einer Einstufung in das niedrigste Niveau des Testinstruments der Rechenfähigkeit an MZP 2 führen.

MZP 3 Die Entwicklung des Niveaumodells des dritten MZP der Rechenfähigkeit der Studie FUNDAMENT erfolgt analog zur Entwicklung des Niveaumodells des vorhergehenden MZP. Es können zwei Grenzen (bei den Punkten 2 und 3 auf der Logit-Skala) identifiziert werden, also eine weniger als am vorherigen MZP. Aus diesen Grenzen lassen sich drei Niveaus ableiten, die mit II, III und IV bezeichnet werden. Das Niveau I, das am vorherigen MZP das niedrigste war, kann an diesem MZP nicht mehr differenziert werden, da keine Items mit Itemschwierigkeiten von $\beta \leq 1$ vorliegen. Insgesamt liegen an diesem MZP 12 Items vor, die alle zur Niveaumodellierung verwendet werden.

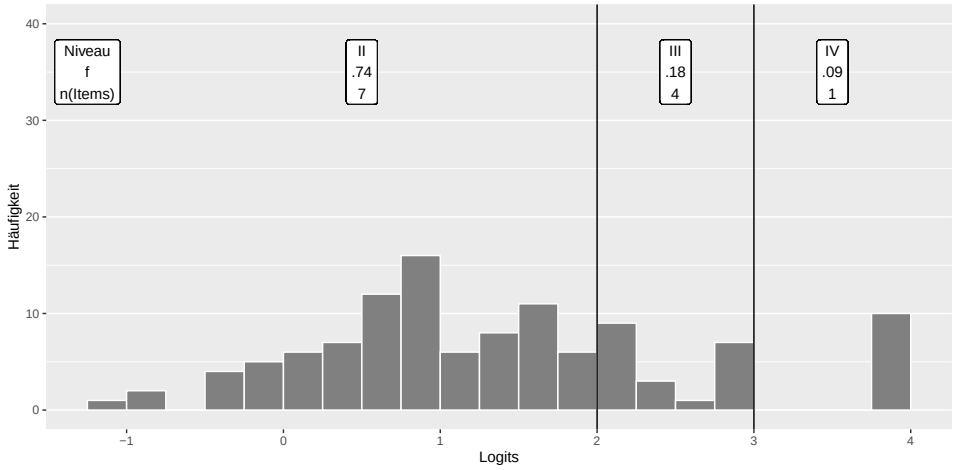
Das entwickelte Niveaumodell ist in Abb. 6.21 abgebildet, es wird die bekannte Darstellungsweise des Histogramms verwendet (gleiche Skalierung der y-Achse und Säulenbreite von 0.25 Logits).

74% der Probanden an diesem MZP sind dem Niveau II zugeordnet. Dies ist ein deutlicher Anstieg gegenüber den 11% an MZP 2, wobei zu berücksichtigen ist, dass es kein Niveau I mehr gibt und somit Niveau II das niedrigste Niveau darstellt. Auffällig sind auch die Steigerungen in den Niveaus III und IV um 14% bzw. 7%.

Die bekannte ›Niveauwanderung‹ wird verwendet, um zu beschreiben, welches Niveau die Probanden an MZP 3 einnehmen, wenn sie zuvor an MZP 2 ein bestimmtes Niveau erreicht haben. Die tabellarische Darstellung findet sich in Tab. 6.35.

Die Ausgangsniveaus (MZP 2) sind in der ersten Spalte der Tab. 6.35 aufgeführt, die weiteren Spalten geben das erreichte Niveau an MZP 3 an.

Abbildung 6.21
Niveaumodell des Testinstruments der Rechenfähigkeit an MZP 3 - FUNDAMENT



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.
Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

Tabelle 6.35
›Niveauwanderung‹ von MZP 2 nach MZP 3 - Testinstrument Rechenfähigkeit (FUNDAMENT)

MZP 2	MZP 3		
Niveau	II	III	IV
I	.89	.09	.02
II	**.67**	.33	
III	.38	**.38**	.25
IV	.25	.25	**.50**

Anmerkungen. MZP = Messzeitpunkt.
Die Abweichungen vom Wert 100 sind auf Rundungsfehler zurückzuführen.

89% der Probanden von Niveau I haben Niveau II erreicht. Dies hängt jedoch damit zusammen, dass es an MZP 3 kein Niveau I mehr gibt. Diese Probanden verbleiben also auf dem niedrigsten Niveau, was auch für die 67% des Ausgangsniveaus II gilt. 33% bzw. 25% der Probanden der Niveau II bzw. III gelingt der Aufstieg um ein Niveau. Die Hälfte der Probanden des höchsten Niveaus kann dieses Niveau halten. Allerdings verschlechtern sich auch 38% der Probanden des dritten Niveaus und 50% der Probanden des höchsten Niveaus.

Im Folgenden wird die inhaltliche Beschreibung anhand der Items der Niveaus vorgenommen:

- Niveau II ($\beta \leq 2$)
  Probanden des niedrigsten Niveaus können sowohl Gleichungen als auch Ungleichungen lösen. Auch die Winkelfunktionen sind ihnen bekannt.

- Niveau III ($2 < \beta \leq 3$)
  Auf diesem Niveau sind Probanden in der Lage, Bruchrechnungen mit Potenzen und Variablen durchzuführen. Außerdem können sie Items zur Analysis (Parabeln und Ableitungen) lösen.

- Niveau IV ($\beta > 3$)
  Im höchsten Niveau beherrschen die Probanden das Rechnen mit Produkten, Potenzen und Wurzeln.

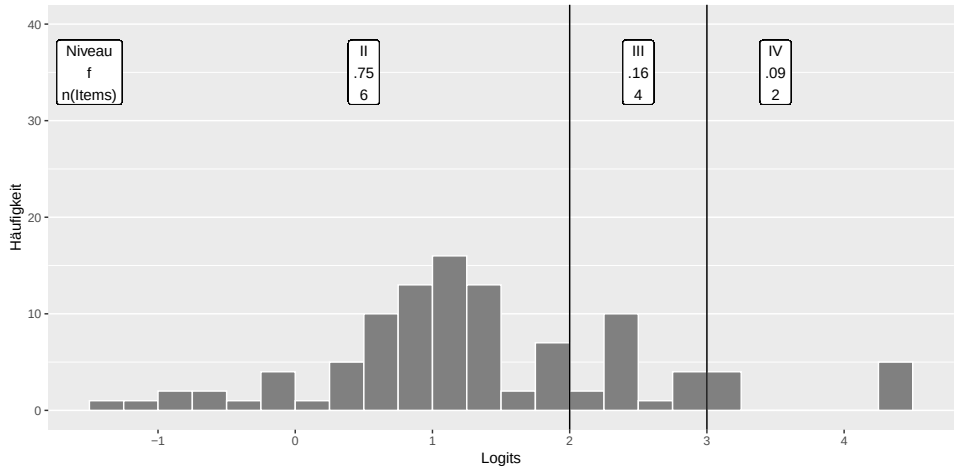
MZP 4 Bei der Niveaumodellierung des vierten MZP des Testinstruments der Rechenfähigkeit der Studie FUNDAMENT konnten ebenfalls zwei Grenzen (bei den Punkten 2 und 3 auf der Logit-Skala) identifiziert werden. Dementsprechend gibt es drei Niveaus, die mit II, III und IV bezeichnet werden. Insgesamt können 12 Items für die Niveaumodellierung verwendet werden, die in Abb. 6.22 dargestellt ist (gleiche Skalierung der y-Achse und Säulenbreite von 0.25 Logits).

Die Verteilung der Probanden ist praktisch identisch mit dem vorhergehenden MZP. Lediglich Niveau II ist um 1% höher, während Niveau III um 2% kleiner ist (Unterschiede im Bereich von Rundungsfehlern).

Um zu prüfen, ob die Probanden von MZP 3 nach MZP 4 andere Niveaus belegen, wird erneut eine Niveauwanderung untersucht, die in

Tab. 6.36 zu finden ist.

Abbildung 6.22
Niveaumodell des Testinstruments der Rechenfähigkeit an MZP 4 - FUNDAMENT



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.

Tabelle 6.36
›Niveauwanderung‹ von MZP 3 nach MZP 4 - Testinstrument Rechenfähigkeit (FUNDAMENT)

MZP 3	MZP 4		
Niveau	II	III	VI
II	**.84**	.09	.06
III	.50	**.30**	.20
IV	.17	.33	**.50**

Anmerkungen. MZP = Messzeitpunkt.
Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

Auf Niveau II verbleiben 84% der Probanden, den restlichen 16% gelingt ein Aufstieg um ein oder sogar zwei Niveaus. Das Niveau III kann von 30% der Probanden gehalten werden, 20% gelingt der Aufstieg in das höchste Niveau, während die Hälfte in das niedrigste Niveau abfällt. Das Niveau IV kann von der Hälfte der Probanden gehalten

werden, ein Drittel wird ein Niveau tiefer eingestuft, 17% sogar zwei Niveaus niedriger.

Die Items dienen erneut der inhaltlichen Beschreibung der Niveaustufen:

- Niveau II ($\beta \leq 2$)
  Auf dem niedrigsten Niveau sind die Probanden in der Lage, Items zu Gleichungen und Ungleichungen zu lösen. Sie können Winkelfunktionen verwenden und Ableitungen von Funktionen bilden.

- Niveau III ($2 < \beta \leq 3$)
  Probanden des dritten Niveaus beherrschen die Multiplikation und können mit Potenzen und Wurzeln rechnen.

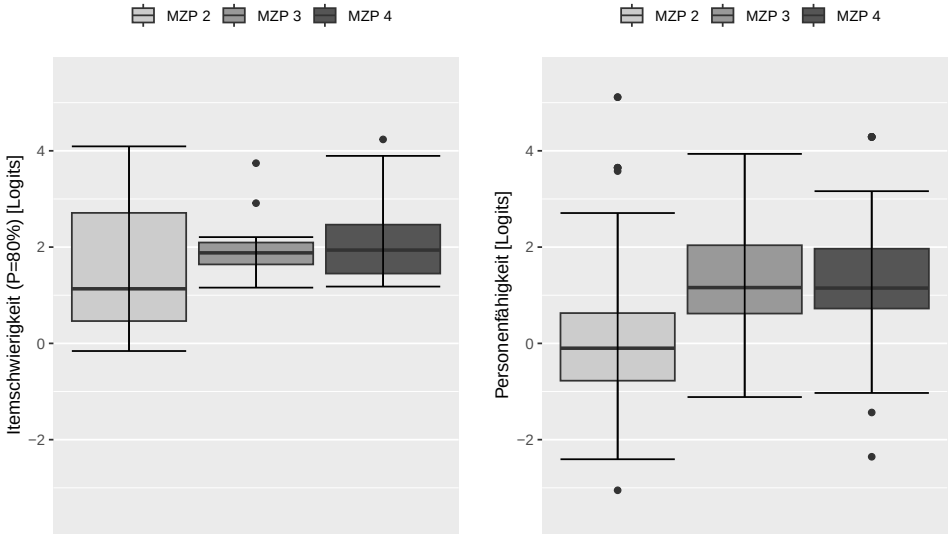
- Niveau IV ($\beta > 3$)
  Die Probanden die dem höchsten Niveau angehören, können Parabeln beschreiben und Logarithmen berechnen.

Vergleich der MZP 2 bis 4 Die drei MZP, an denen die Testinstrumente der Rechenfähigkeit der Studie FUNDAMENT eingesetzt wurden, werden ebenfalls miteinander verglichen. Die nach dem 2pl-Birnbaum-Modell berechneten Itemschwierigkeiten ($P(u_{ij} = 1) = 80\%$) und Personenfähigkeiten sind in Abb. 6.23 zu finden.

Wird die Itemschwierigkeit ($P(u_{ij} = 1) = 80\%$) betrachtet, so zeigt das Diagramm, dass der Median der Itemschwierigkeit für MZP 2 knapp über 1 liegt, während die Itemschwierigkeiten für die MZP 2 und 3 fast 2 betragen. Die Größe des Kastens nimmt von MZP 2 nach MZP 3 stark ab, von MZP 3 nach MZP 4 wird der Kasten wieder etwas größer. Ein kleiner Kasten bedeutet eine geringe Streuung innerhalb der Quartile 1 und 3 (siehe Kap. 6.1.1). Ob es statistisch signifikante Unterschiede zwischen den drei MZP gibt, wird mit der einfaktoriellen ANOVA untersucht. Die Voraussetzungen zur Berechnung einer ANOVA sind erfüllt. Es zeigen sich keine signifikanten Unterschiede der Itemschwierigkeiten zwischen den drei MZP ($F(2, 41) = 1.09$, $p = .345$).

Ein Vergleich der Box-Whisker-Diagramme der Itemschwierigkeiten und der Personenfähigkeiten zeigt, dass die Itemschwierigkeiten aufgrund

Abbildung 6.23
Box-Whisker-Diagramm der Itemschwierigkeit ($P(u_{ij} = 1) = 80\%$) und der Personenfähigkeit der Testinstrumente der Rechenfähigkeit (FUNDAMENT) an den MZP 2 bis 4



Anmerkungen. MZP = Messzeitpunkt.

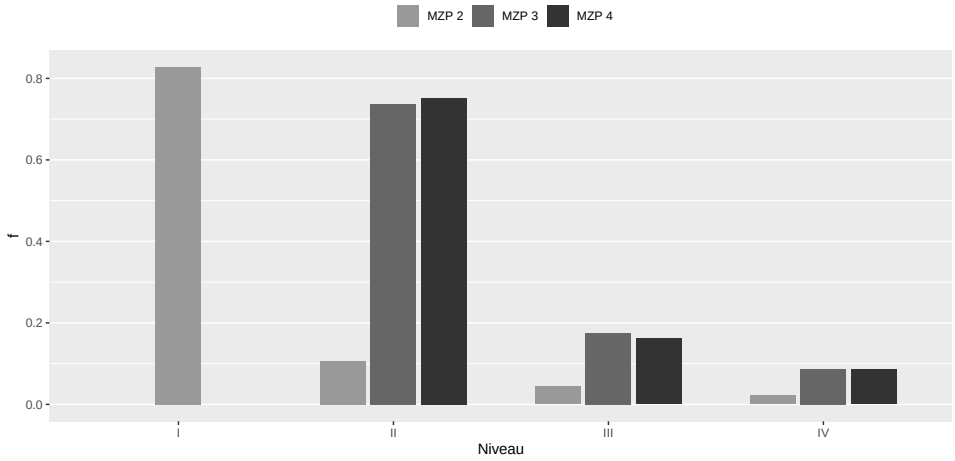
der Lösungswahrscheinlichkeit von 80% um ca. einen Logit in positive Y-Richtung verschoben sind. Bei einer Lösungswahrscheinlichkeit von $P(u_{ij} = 1) = 50\%$ liegen die mittleren Itemschwierigkeiten zwischen -0.05 und 0.92 (siehe Kap. 6.2.3.2).

Die Box-Whisker-Diagramme der Personenfähigkeit (Abb. 6.23) zeigen ein ähnliches Bild wie die Testinstrumente der Technischen Mechanik. Der Median liegt an MZP 2 knapp unter 0, während für die MZP 3 und 4 der Median der Personenfähigkeit über 1 liegt. Ein signifikanter Unterschied zwischen den MZP wird durch die berechnete ANOVA bestätigt ($F(2, 490) = 71.13$, $p < .001$). Zwischen den MZP 2 und MZP 3 ($p < .001$, $d = -0.39$) sowie den MZP 2 und MZP 4 ($p < .001$, $d = -0.47$) gibt es signifikante Unterschiede mit einem kleinen Effekt. Zwischen den MZP 3 und 4 gibt es keine signifikanten Unterschiede.

Es gibt also keine signifikanten Unterschiede zwischen den MZP bezüglich der Itemschwierigkeiten. Die Personenfähigkeiten unterscheiden sich jedoch signifikant (MZP 2 und die beiden weiteren MZP). Diese Aussagen lassen sich auch durch die Kompetenzniveaus stützen, die für

alle drei MZP in Abb. 6.24 als Säulendiagramm dargestellt sind.

Abbildung 6.24
Verteilung der Kompetenzniveaus der Testinstrumente der Rechenfähigkeit (FUNDAMENT) an den MZP 2 bis 4



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit.

Die meisten Probanden befinden sich jeweils auf den niedrigsten Niveaus (MZP 2: Niveau I, MZP 3 und 4: Niveau II). Außerdem ist ein Anstieg von MZP 2 zu MZP 3 und 4 auf allen Kompetenzniveaus zu beobachten. Der Unterschied zwischen MZP 3 und 4 ist minimal. Die statistische Überprüfung der genannten Unterschiede erfolgt mit dem Kruskal-Wallis-Test, der signifikante Unterschiede zwischen den MZP bestätigt ($H(2) = 281.61, p < .001$). In der Post-hoc-Berechnung können signifikante Unterschiede mit großer Effektstärke zwischen MZP 2 und 3 ($p < .001$, $r = 0.72$) sowie MZP 2 und 4 ($p < .001$, $r = 0.71$) nachgewiesen werden. Durch den Wegfall von Kompetenzniveau 1 sind diese signifikanten Unterschiede zu erwarten. Der Unterschied zwischen den MZP 3 und 4 ist nicht signifikant.

Der Anstieg von MZP 2 nach MZP 3 kann bei den Testinstrumenten der Technischen Mechanik damit begründet werden, dass die Probanden noch keine Vorlesungen zu diesem Thema besucht haben. Daher müssen sie sich die Kenntnisse hauptsächlich über die Schulphysik angeeignet haben, sofern sie in der Sekundarstufe II einen Kurs besucht haben. Diese Begründung kann nicht auf die Rechenfähigkeit übertra-

gen werden. Die Testinstrumente enthalten Items, die grundlegende Rechenfähigkeiten abfragen, die bereits in der Schule erworben werden müssen. In den Veranstaltungen der Mathematik in den ersten beiden Fachsemestern werden diese Inhalte in der Regel nicht wiederholt. Dementsprechend müssen die Probanden diese Kenntnisse selbstständig oder z.B. in Tutorien wiederholen und sich so aneignen. Da auch die Itemschwierigkeiten keine signifikanten Unterschiede aufweisen, ist dies die plausibelste Erklärung. Dass es keine signifikanten Unterschiede zwischen MZP 3 und 4 gibt, kann genau damit erklärt werden. Die Probanden scheinen zwar kein neues Wissen in diesem Bereich erworben zu haben, aber letztendlich reicht ihr vorhandenes Wissen aus, um das Niveau zu halten. Ein möglicher Wissenserwerb reicht jedoch nicht aus, um ein höheres Kompetenzniveau zu erreichen.

6.3.3 Rechenfähigkeit (ALSTER)

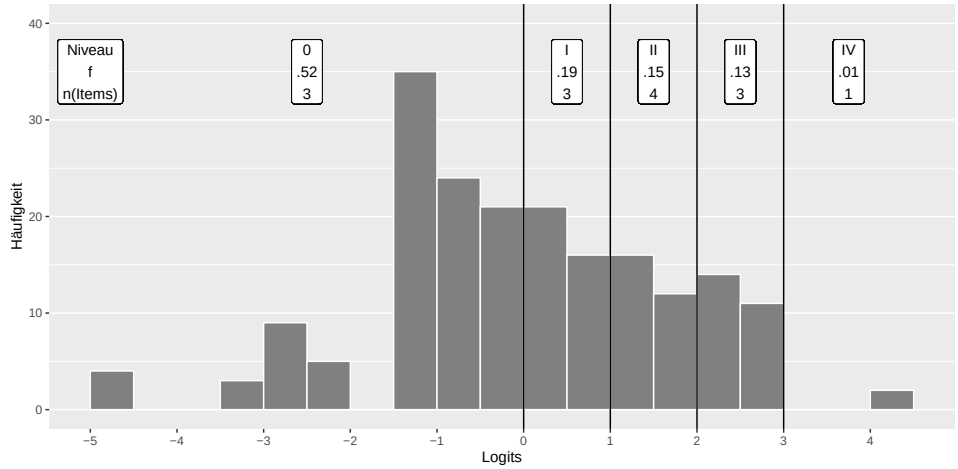
Lediglich für die Testinstrumente der Rechenfähigkeit der Studie ALSTER zeigt das 1pl-Rasch-Modell die beste Passung bei der IRT-Skalierung (siehe Kap. 6.2.3.3). Demnach könnten die Itemschwierigkeiten mit einer Lösungswahrscheinlichkeit von 50% für die Niveaumodellierung verwendet werden. Da die Testinstrumente zur Messung der Rechenfähigkeit von FUNDAMENT und ALSTER nicht gemeinsam skaliert werden dürfen (siehe Kap. 6.2.1), wurde eine getrennte Skalierung durchgeführt. Ein direkter Vergleich der entsprechenden Ergebnisse ist daher nicht zulässig. Um jedoch zumindest tendenzielle Aussagen treffen zu können, wird auch bei dieser Niveaumodellierung die Lösungswahrscheinlichkeit von 80% verwendet. Eine Überprüfung der Rangfolge der Items ist nicht erforderlich, da aufgrund der spezifischen Objektivität der Items im 1pl-Rasch-Modell die Rangfolge der Items erhalten bleibt.

MZP 2 Die Niveaumodellierung der Testinstrumente der Rechenfähigkeit der Studie ALSTER erfolgt analog zu der zuvor durchgeführten Niveaumodellierung. Der einzige Unterschied besteht darin, dass die vorliegenden Daten nicht mit dem 2pl-Birnbaum-Modell IRT, sondern mit dem 1pl-Rasch-Modell skaliert wurden. Dennoch werden die Itemschwierigkeiten mit Lösungswahrscheinlichkeiten von $P(u_{ij} = 1) = 80\%$

verwendet.

An MZP 2 wurden vier Niveaugrenzen identifiziert, die bei den Punkten 0, 1, 2 und 3 auf der Logit-Skala liegen. Diese Niveaugrenzen führen zu fünf Kompetenzniveaus, die mit 0, I, II, III und IV bezeichnet werden. So umfasst Niveau 0 Itemschwierigkeiten (bei $P(u_{ij} = 1) = 80\%$) von $\beta \leq 0$, Niveau I umfasst $0 < \beta \leq 1$ usw. Alle 16 Items werden für die Niveaumodellierung verwendet. Das Niveaumodell für den zweiten MZP wird in einem Histogramm dargestellt (Abb. 6.25) wobei im Gegensatz zu den vorherigen Histogrammen eine Säulenbreite von 0.5 Logits verwendet wird.

Abbildung 6.25
Niveaumodell des Testinstruments der Rechenfähigkeit an MZP 2 - ALSTER



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.

Aufgrund der hoch angesetzten Lösungswahrscheinlichkeit ($P(u_{ij} = 1) = 80\%$) zeigt sich erneut die Problematik, dass die Niveaus I, II und III jeweils einen Bereich von einem Logit abdecken, Niveau 0 jedoch einen Bereich von 5 Logits. Weitere Abgrenzungen können im Niveau 0 nicht vorgenommen werden, da hierfür keine Items mit entsprechenden Itemschwierigkeiten vorliegen, sodass eine inhaltliche Beschreibung nicht möglich ist. In den höheren Niveaus können die Fähigkeiten der Probanden also besser differenziert werden als in Niveau 0.

Auch diese Verteilung ist rechtsschief, knapp die Hälfte (52%) der

Probanden befindet sich in Niveau 0. Die restlichen Probanden verteilen sich abnehmend auf die Niveaus I bis III (19% bis 13%), Niveau IV umfasst nur zwei Probanden.

Die inhaltliche Beschreibung der Niveaus erfolgt wiederum durch die entsprechenden Items. Es gibt nur zwischen einem und vier Items pro Niveau. Die Inhalte werden nur einmal genannt, können aber in höheren Niveaus mit einer erhöhter Komplexität erneut vorkommen, ohne dass dies noch einmal explizit erwähnt wird.

- Niveau 0 ($\beta \leq 0$)
  In dem niedrigsten Niveau sind die Probanden in der Lage, Summen von Brüchen sowie Summen und Differenzen von Vektoren zu berechnen.

- Niveau I ($0 < \beta \leq 1$)
  Probanden können Skalarprodukte von Vektoren und Stammfunktionen bestimmen. Winkelfunktionen können ebenfalls reproduziert werden.

- Niveau II ($1 < \beta \leq 2$)
  Items zu Gleichungen können von den Probanden gelöst werden.

- Niveau III ($2 < \beta \leq 3$)
  In der Analysis können Probanden Ableitungen bilden.

- Niveau IV ($\beta > 3$)
  In dem höchsten Niveau sind die Probanden in der Lage, Matrizen zu multiplizieren.

Die Inhalte der Niveaus sind durch den Kernlehrplan Mathematik für die Sekundarstufe II für Gymnasien und Gesamtschulen in NRW (MSB NRW, 2023) abgedeckt. Es kann geprüft werden, ob es auch bei ALSTER signifikante Unterschiede innerhalb der Niveaus gibt, je nachdem ob in der Sekundarstufe II ein Mathematik Grund- oder Leistungskurs belegt wurde. Die Wahl des Mathematik-Kurses hat dabei einen signifikanten Einfluss auf das Kompetenzniveau mit einem mittleren Effekt ($H(1) = 28.78$, $p < .001$, $|r| = 0.39$). Probanden, die in der Schule einen Grundkurs besucht haben, erreichen an MZP 2

niedrigere Kompetenzniveaus als Probanden, die in der Sekundarstufe II einen Leistungskurs besucht haben. Ein möglicher Einfluss der Mathematik-Abschlussnote kann nicht überprüft werden, da diese Angabe im Datensatz von ALSTER nicht vorliegt.

Weitere demographische Variablen (Alter, Muttersprache, Bildungsherkunft, Art der Hochschulzugangsberechtigung und Note der HZB) wurden hinsichtlich ihres Einflusses auf das erreichte Kompetenzniveau der Rechenfähigkeit an MZP 2 untersucht. Da nur zwei Probanden in Niveau IV vorhanden sind, wird diese Niveau in den Berechnungen nicht berücksichtigt. Der Kruskal-Wallis-Test zeigt für die Variablen Art des Erwerbs der Hochschulzugangsberechtigung ($H(4) = 11.55$, $p = .021$), Muttersprache ($H(4) = 10.838$, $p = .028$) und Bildungsherkunft ($H(4) = 18.26$, $p < .001$) signifikante Unterschiede zwischen den Kompetenzniveaus. Die entsprechenden Post-hoc-Tests zeigen jedoch keine signifikanten Unterschiede oder lassen keine sinnvolle Interpretation zu.

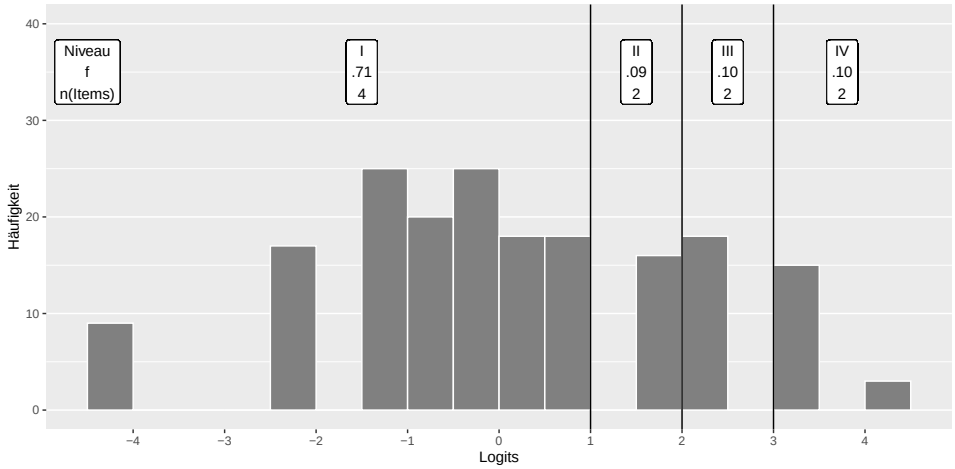
Für die HZB-Note gibt es signifikante Unterschiede zwischen den Niveaus ($F(4, 187) = 8.44$, $p < .001$). Diese signifikanten Unterschiede zeigen sich zwischen Niveau 0 und den Niveaus II ($p < .001$, $d = .89$) und III ($p < .001$, $d = .99$) mit einer hohen Effektstärke, sowie zwischen Niveau I und den Niveaus II ($p < .001$, $d = .70$) und III ($p < .001$, $d = .82$) mit einer mittleren bis hohen Effektstärke. Damit lässt sich die Aussage, dass bessere HZB-Noten zu einem höheren Kompetenzniveau bei Studienbeginn führen, auch für das Testinstrument der Rechenfähigkeit der Studie ALSTER replizieren.

MZP 3 Für den dritten MZP des Testinstruments der Rechenfähigkeit der Studie ALSTER lassen sich drei Grenzen (bei den Punkten 1, 2 und 3 auf der Logit-Skala) und daraus resultierend vier Niveaus identifizieren. Die Bezeichnung der Grenzen erfolgt nach dem bekannten Schema, sodass die Niveaus als I, II, III und IV bezeichnet werden. Damit gibt es eine Grenze und dementsprechend ein Niveau weniger, als beim vorherigen MZP. Dies liegt daran, dass es keine Items gibt, die eine Itemschwierigkeit von $\beta \leq 0$ haben. Alle zehn eingesetzten Items können für die Niveaumodellierung verwendet werden.

In Abb. 6.26 ist das für den dritten MZP entwickelte Niveaumodell

abgebildet. Das Histogramm folgt den gleichen Vorgaben wie das im vorhergehenden MZP.

Abbildung 6.26
Niveaumodell des Testinstruments der Rechenfähigkeit an MZP 3 - ALSTER



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit; n(Items) = Anzahl der Items.

An diesem MZP sind 71% der Probanden dem niedrigsten Niveau zugeordnet. Dies ist der gleiche Wert, wenn man die beiden niedrigsten Niveaus (0 und I) am vorherigen MZP zusammenfasst. Außerdem sind die Niveaus II und III um 6% bzw. 3% zurückgegangen, während das höchste Niveau um 9% zugewonnen hat.

Eine ›Niveauwanderung‹ soll zeigen, in welche Niveaus sich die Probanden, die an MZP 2 ein bestimmtes Niveau erreicht haben, zu MZP 3 einordnen. In Tab. 6.37 ist die tabellarische Übersicht zu finden.

Wie bei den vorherigen ›Niveauwanderungen‹ ist das Ausgangsniveau (MZP 2) in der ersten Spalte aufgeführt, die an MZP 3 erreichten Niveaus sind in den anderen Spalten zu finden. 95% der Probanden des niedrigsten Niveaus erreichen an MZP 3 das Niveau I. Da dies nun das niedrigste Niveau ist, verbleiben sie auf dem niedrigsten Niveau. Auch 78% der Probanden des Niveaus I erreichen keine Verbesserung und verbleiben somit in dem niedrigsten Niveau. In Niveau II verbleiben 21% der Probanden, während sich 29% um zwei Niveaus in das höchste Niveau verbessern. 48% können das Niveau II halten, es verschlechtern

Tabelle 6.37
›Niveauwanderung‹ von MZP 2 nach MZP 3 - Testinstrument Rechenfähigkeit (ALSTER)

MZP 2	MZP 3			
Niveau	I	II	III	VI
0	.95	.05		
I	**.78**	.08	.05	.08
II	.39	**.21**	.11	.29
III	.20	.08	**.48**	.24
IV			.50	**.50**

Anmerkungen. MZP = Messzeitpunkt.
Die Abweichung vom Wert 100 ist auf Rundungsfehler zurückzuführen.

sich 20% sogar um zwei Niveaus auf das niedrigste Niveau. Nur zwei Probanden befinden sich an MZP 2 in Niveau IV, während einer das Niveau halten kann, verschlechtert sich der andere um ein Niveau.

Nachfolgend werden die Niveaus anhand der Items inhaltlich beschrieben:

- Niveau I ($\beta \leq 1$)
  In diesem Niveau sind Probanden in der Lage, Gleichungen zu lösen und Winkelfunktionen anzuwenden.

- Niveau II ($1 < \beta \leq 2$)
  Probanden können mit Potenzen und Variablen in Brüchen rechnen.

- Niveau III ($2 < \beta \leq 3$)
  Probanden in diesem Niveau können Ableitungen in der Analysis bilden.

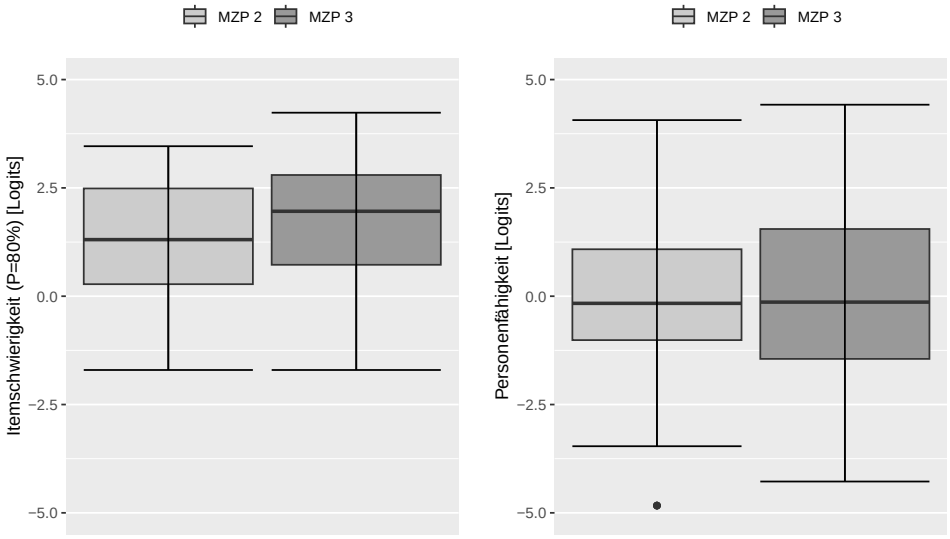
- Niveau IV ($\beta > 3$)
  Das höchste Niveau umfasst das Rechnen mit Produkten, Potenzen und Wurzeln. Außerdem sind Probanden mit Parabeln vertraut.

Da aufgrund von Kodierungsproblemen die Daten des Testinstruments der Rechenfähigkeit des vierten MZP von ALSTER nicht verwen-

det werden können, wird für ALSTER keine weitere Niveaumodellierung durchgeführt.

Vergleich der MZP 2 und 3 Ein Vergleich der beiden MZP der Testinstrumente der Rechenfähigkeit der Studie ALSTER erfolgt zunächst anhand der nach dem 1pl-Rasch-Modell berechneten Itemschwierigkeiten ($P(u_{ij} = 1) = 80\%$) und Personenfähigkeiten. Entsprechende Box-Whisker-Diagramme der Itemschwierigkeiten und Personenfähigkeiten finden sich in Abb. 6.27.

Abbildung 6.27
Box-Whisker-Diagramm der Itemschwierigkeit ($P(u_{ij} = 1) = 80\%$) und der Personenfähigkeit der Testinstrumente der Rechenfähigkeit (ALSTER) an den MZP 2 und 3



Anmerkungen. MZP = Messzeitpunkt.

Wie aus Abb. 6.27 hervorgeht, ist der Median der Itemschwierigkeit an MZP 3 höher als der von MZP 2. Der Kasten und damit die Streuung der Werte innerhalb der Quartile 1 und 3 ist an MZP 2 ausgeprägter. Die statistische Überprüfung erfolgt mittels t-Test für unabhängige Stichproben. Die Anwendung des t-Tests für unabhängige Stichproben ist an mehrere Voraussetzungen geknüpft (Bortz & Schuster, 2010; Eid et al., 2017; Field et al., 2012):

1. voneinander unabhängige Messwerte zwischen und innerhalb der Stichproben

2. Normalverteilung innerhalb beider Stichproben

3. Homoskedastizität (homogene Varianzen innerhalb der Stichproben) (Bortz & Schuster, 2010; Eid et al., 2017; Field et al., 2012)

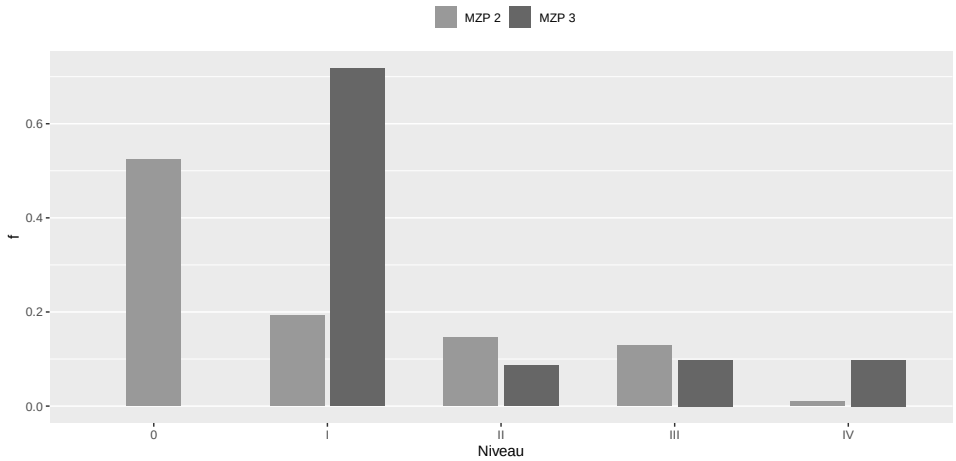
Diese Bedingungen sind im vorliegenden Fall erfüllt. Der t-Test wird mit der Funktion ›t.test‹ aus dem ›R‹-Paket ›stats‹ (R Core Team, 2023) berechnet. Der t-Test kann den in Abb. 6.27 erkennbaren Unterschied nicht bestätigen, das Ergebnis ist nicht signifikant ($t(22) = -0.85$, $p = .202$). Aufgrund der Lösungswahrscheinlichkeit von 80% sind die Itemschwierigkeiten in positive Y-Richtung verschoben. Die mittleren Itemschwierigkeiten betragen -0.24 und 0.34 bei einer Lösungswahrscheinlichkeit von $P(u_{ij} = 1) = 50\%$.

Die Box-Whisker-Diagramme der Personenfähigkeiten zeigen nahezu identische Mediane, lediglich die Streuung ist beim dritten MZP größer. Auch für die Personenfähigkeit wird ein t-Test durchgeführt, der keinen signifikanten Unterschied zwischen den beiden MZP zeigt ($t(375) = -0.49$, $p = .311$).

Somit gibt es weder zwischen den Itemschwierigkeiten noch zwischen den Personenfähigkeiten an den beiden MZP signifikante Unterschiede. Betrachtet man hingegen die erreichten Kompetenzniveaus, so zeigen sich Unterschiede zwischen den beiden MZP, wie aus Abb. 6.28 hervorgeht.

Sowohl an MZP 2 als auch an MZP 3 sind die meisten Probanden dem niedrigsten Niveau (MZP 2: Niveau 0, MZP 3: Niveau I) zugeordnet. Während die Niveaus II und III bei MZP 2 stärker ausgeprägter sind, ist das höchste Niveau an MZP 3 häufiger vertreten. Mit einem Chi-Quadrat-Test können diese Beobachtungen statistisch überprüft werden, da jedoch nicht alle Niveaus mindestens fünf Probanden aufweisen, sollte stattdessen der Fisher-Yates-Test (auch exakter Fisher-Test genannt) verwendet werden (Eid et al., 2017). Die Berechnung erfolgt mit dem ›R‹-Paket ›stats‹ (R Core Team, 2023) und der Funktion ›fisher.test‹. Die Effektstärke wird mit der Funktion ›effectsize‹ aus dem gleichnamigen ›R‹-Paket (Ben-Shachar et al., 2020) berechnet. Der

Abbildung 6.28
Verteilung der Kompetenzniveaus des Testinstruments der Rechenfähigkeit (ALSTER) an den MZP 2 und 3



Anmerkungen. MZP = Messzeitpunkt; f = relative Häufigkeit.

Fisher-Yates-Test zeigt signifikante Unterschiede zwischen den beiden MZP mit großer Effektstärke ($p < .001$, $V = 0.67$). Dieses Ergebnis ist nicht überraschend, da sich allein durch den Wegfall des Kompetenzniveaus 0 eine deutliche Verschiebung der Verteilung ergibt.

Wie anhand Tab. 6.37 gezeigt werden kann, ist es jeweils über 20% der Probanden gelungen von Niveau II oder III am dritten MZP das Niveau IV und damit das höchste Niveau zu erreichen. Dementsprechend sind die Niveaus II und III weniger stark besetzt, da aus den beiden unteren Niveaus nur wenige Probanden den Sprung auf ein höheres Kompetenzniveau schaffen (siehe Tab. 6.37).

Vergleich Rechenfähigkeit - FUNDAMENT und ALSTER Wie bereits erwähnt, ist ein direkter Vergleich der Ergebnisse der Testinstrumente der Rechenfähigkeit zwischen den beiden Studien nicht möglich. Es lassen sich jedoch Tendenzen zwischen MZP 2 und 3 erkennen. Hinsichtlich des erreichten Niveaus an MZP 2 erweist es sich in beiden Studien als signifikant, ob die Probanden in der Schule einen Grund- oder Leistungskurs Mathematik besucht haben. Auch die Note der Hochschulzugangsberechtigung hat einen signifikanten Einfluss auf das

zu Studienbeginn erreichte Niveau. In FUNDAMENT erweist sich auch die letzte Fachnote Mathematik als signifikant (für ALSTER liegen keine entsprechenden Daten vor). Sowohl die Kurswahl Mathematik in der Sekundarstufe II und die letzte Fachnote Mathematik als auch die Note der Hochschulzugangsberechtigung sind somit prädiktiv für das erreichte Niveau an MZP 2.

In FUNDAMENT ist es den Probanden gelungen, von einem mittleren Niveau auf ein höheres oder sogar auf das höchste Niveau am dritten MZP aufzusteigen. Dies zeigt sich auch in ALSTER. So kann in beiden Studien gezeigt werden, dass Probanden von MZP 2 (mittleres Niveau) bis MZP 3 Wissen erwerben oder zumindest in der Lage sind, die notwendigen Items zu beantworten, um das höchste Niveau zu erreichen. Es zeigt sich aber auch, dass die Probanden auf den niedrigsten Niveaus als Risikogruppe bezeichnet werden müssen, da es ihnen in der Regel nicht gelingt, an weiteren MZP ein höheres Niveau zu erreichen.

Damit ist die Entwicklung, Beschreibung und Interpretation der Niveaumodellierung abgeschlossen. In den folgenden Kapiteln werden die Forschungsfragen auf der Grundlage der Niveaumodelle beantwortet.

6.4 Prüfung der Hypothesen

In diesem Kapitel werden die in den Kap. 3 vorgestellten Forschungsfragen beantwortet und die Ergebnisse interpretiert. Dabei wird jeweils auf die postulierten Hypothesen Bezug genommen.

6.4.1 Forschungsfrage 1

Mit ersten Forschungsfrage (F1) soll geklärt werden, ob die Befunde von ALSTER in FUNDAMENT repliziert werden können (siehe Kap. 3). Hypothese H1.1 besagt, dass es für den Inhaltsbereich Technische Mechanik keine signifikanten Unterschiede zwischen den beiden Studien an den MZP 2 bis 4 gibt.

Die Frage, ob es signifikante Unterschiede zwischen den Studien gibt, wird anhand der berechneten Personenfähigkeiten und der erreichten Niveaus der Probanden geklärt. Zeigen die Personenfähigkeiten an allen drei MZP keine signifikanten Unterschiede zwischen den beiden Studien,

so gelten die Ergebnisse von ALSTER als repliziert. Als zusätzliche Prüfung werden die erreichten Kompetenzniveaus an den drei MZP auf signifikante Unterschiede geprüft. Da die Kompetenzniveaus der Probanden aus deren Personenfähigkeiten abgeleitet werden, ist zu erwarten, dass die Prüfung der Niveaus zu dem gleichen Ergebnis führt.

Um festzustellen, ob es signifikante Unterschiede zwischen den Personenfähigkeiten gibt, wird ein unabhängiger t-Test berechnet (siehe Kap. 6.3.3). Die Varianzhomogenität wird mit dem Levene-Test geprüft. Dieser wird mit der Funktion ›leveneTest‹ aus dem ›R‹-Paket ›car‹ (Fox & Weisberg, 2019) berechnet. Ein signifikantes Ergebnis des Levene-Tests würde bedeuten, dass keine Varianzhomogenität vorliegt und anstelle des t-Tests der Welch-Test berechnet werden sollte (Eid et al., 2017). Der Welch-Test wird mit der gleichen Funktion (›t.test‹) wie der t-Test berechnet. Die Berechnung der Effektstärke Cohen's *d* erfolgt mit der Funktion ›effectsize‹ aus dem gleichnamigen ›R‹-Paket (Ben-Shachar et al., 2020).

Die Ergebnisse des t-Tests sind in Tab. 6.38 aufgelistet. Für MZP 2 wurde der Welch-Test aufgrund von Varianzheterogenität berechnet.

Tabelle 6.38
Ergebnisse des t-Tests für unabhängige Stichproben der Personenfähigkeiten (FUNDAMENT und ALSTER)

	t	*df*	*p*	*d*
MZP 2[a]	−8.91	316	<.001	−0.86
MZP 3	7.24	297	<.001	0.86
MZP 4	2.67	178	.008	0.40

Anmerkungen. t = t-Wert; df = Freiheitsgrade; p = Wahrscheinlichkeit; d = Cohen's d; MZP = Messzeitpunkt.
[a] Aufgrund von Varianzheterogenität wurde der Welch-Test berechnet.

Für alle drei MZP zeigen die durchgeführten Tests signifikante Ergebnisse. Demzufolge unterscheiden sich die Mittelwerte der beiden Stichproben und es kann nicht davon ausgegangen werden, dass die Personenfähigkeiten der beiden Studien gleich sind. Bei MZP 2 und 3 sind die Effektstärken hoch, während bei MZP 4 nur von einem geringen Effekt gesprochen werden kann.

Auffällig sind die Vorzeichen der Effektstärken. Während der zweite MZP ein negatives Vorzeichen hat, haben die beiden anderen MZP ein positives Vorzeichen. Für MZP 2 bedeutet dies, dass der Mittelwert der Studie FUNDAMENT kleiner ist als der der Studie ALSTER. Bei MZP 2 und 3 sind die Mittelwerte von ALSTER niedriger.

Es ist zu berücksichtigen, dass an MZP 2 viele Werte im Datensatz fehlen. So gibt es bei MZP 2 für ALSTER 1088 fehlende Werte (7%), während es an dem gleichen MZP für FUNDAMENT nur 343 (3%) sind (siehe Tab. 6.2). Dies führte zur Imputation vieler fehlender Werte, entsprechend stieg der Mittelwert des Summenscores um 2.6, während er nach der Imputation bei FUNDAMENT nur um 0.5 stieg (siehe Tab. 6.3). Dies könnte ein möglicher Grund dafür sein, dass an MZP 2 ALSTER einen höheren Mittelwert der Personenfähigkeit hat.

Die Überprüfung der Kompetenzniveaus auf signifikante Unterschiede zwischen den beiden Studien erfolgt mit dem Fisher-Yates-Test (siehe Kap. 6.3.3). Für den zweiten MZP liegen signifikante Unterschiede zwischen den Niveaus der beiden Studien vor ($p < .001$, $V = 0.35$). Dies gilt auch für MZP 3 ($p < .001$, $V = 0.36$). Allerdings liefert der Fisher-Yates-Test für den vierten MZP kein signifikantes Ergebnis ($p = .100$).

Wie zu erwarten, sind weitestgehend die Ergebnisse zwischen den Personenfähigkeiten und den Kompetenzniveaus gleich. Eine Abweichung gibt es am vierten MZP, während die Personenfähigkeiten signifikante Unterschiede (kleiner Effekt) aufweisen, zeigen die Kompetenzniveaus dies nicht. Der beobachtbare kleine Effekt der Personenfähigkeiten wird also durch die Niveaumodellierung kompensiert.

Im Ergebnis zeigt sich jedoch, dass es signifikante Unterschiede in den Befunden der Testinstrumente der Technischen Mechanik zwischen FUNDAMENT und ALSTER gibt. Somit muss H1.1 falsifiziert werden. Die Gründe für diese signifikanten Unterschiede können nicht abschließend geklärt werden. Wie bereits in Kap. 5.5 dargelegt, unterscheiden sich die Stichproben der beiden Studien in verschiedenen demographischen Variablen (mit einem kleiner Effekt). Diese Abweichungen können zu den oben genannten Auswirkungen führen.

Aus H1.2 geht hervor, dass die unterschiedlichen Antwortformate der Testinstrumente der Rechenfähigkeit keinen Einfluss auf die Rang-

folge der Itemschwierigkeiten haben. Da in ALSTER ein offenes und in FUNDAMENT ein geschlossenes Antwortformat verwendet wurde (siehe Kap. 5.2), kann anhand der Rangfolge der Itemschwierigkeiten geprüft werden, ob eine Skalierung als gemeinsamer Datensatz zulässig ist. Wenn die Items die gleiche Rangfolge aufweisen, können beide Studien gemeinsam skaliert werden (siehe Kap. 6.2.1).

Bereits in Kap. 6.2.1 wurde die Rangfolge der Itemschwierigkeiten mit Hilfe der Pearson-Produkt-Moment-Korrelation überprüft. Eine hohe signifikante Korrelation weist auf eine gleiche Rangfolge der Itemschwierigkeiten hin. Da jedoch keine signifikante Korrelation zwischen den Itemschwierigkeiten der Testinstrumente der Rechenfähigkeit an den MZP 2 und 3 (FUNDAMENT und ALSTER) besteht (siehe Tab. 6.5), muss auch H1.2 falsifiziert werden. In den betrachteten Stichproben gibt es einen signifikanten Unterschied zwischen den beiden Antwortformaten.

Da beide Hypothesen nicht bestätigt werden konnten muss auch die Forschungsfrage, ob die Ergebnisse von ALSTER für die beiden Inhaltsbereiche durch die Studie FUNDAMENT repliziert werden können, verneint werden.

6.4.2 Forschungsfrage 2

Ob signifikante Zusammenhänge zwischen den modellierten Kompetenzniveaus der Technischen Mechanik bzw. der Mathematik (FUNDAMENT und ALSTER) und dem Studienerfolg bestehen, soll in der zweiten Forschungsfrage geklärt werden (siehe Kap. 3). Sollten diese signifikanten Korrelationen vorliegen, dann können die Kompetenzniveaus als Prädiktoren für den Studienerfolg angesehen werden. Als Studienerfolg wird hier die erreichte Klausurnote in der Technischen Mechanik bzw. Mathematik am Ende des ersten bzw. zweiten Fachsemesters betrachtet.

Die statistische Überprüfung auf signifikante Zusammenhänge, erfolgt mittels der Spearman-Korrelation. Dazu werden die Daten in Rangwerte transformiert und anschließend die Produkt-Moment-Korrelation nach Pearson berechnet (Bortz & Schuster, 2010; Eid et al., 2017; Field et al., 2012). Die Berechnung der Spearman-Korrelation erfolgt durch

die Funktion ›cor.test‹ mit der Option ›method = 'spearman'‹ aus dem ›R‹-Paket »stats« (R Core Team, 2023). Als Effektstärke für die Spearman-Korrelation wird der Spearman-Korrelationskoeffizient (r_s) verwendet (Bortz & Schuster, 2010; Eid et al., 2017; Field et al., 2012).

Für r_s gilt die gleiche Interpretation der Effektstärke wie für den Pearson Korrelationskoeffizienten (r):

- $|r_s| \geq 0.1$ - kleiner Effekt
- $|r_s| \geq 0.3$ - mittlerer Effekt
- $|r_s| \geq 0.5$ - großer Effekt (Cohen, 1988).

Die Effektstärke wird direkt bei der Berechnung der Spearman-Korrelation ausgegeben. In Tab. 6.39 sind die Spearman-Korrelationskoeffizienten aufgelistet.

Tabelle 6.39
Spearman Korrelationskoeffizienten zwischen erreichten Kompetenzniveaus (Technische Mechanik) und den Klausurnoten

Niveau	TM 1	TM 2
MZP 2	$-.18^{***}$	$-.24^{***}$
MZP 3	$-.69^{***}$	$-.59^{***}$
MZP 4	$-.44^{***}$	$-.51^{***}$

Anmerkungen. TM 1 = ›Technische Mechanik 1‹; TM 2 = ›Technische Mechanik 2‹; MZP = Messzeitpunkt.
$^{***}p < .001$

Für alle MZP bestehen hochsignifikante Zusammenhänge zwischen den erreichten Kompetenzniveaus der Technischen Mechanik und den erzielten Klausurnoten. Während MZP 2 geringe negative Effekte aufweist, sind es bei MZP 3 hohe negative Effektstärken und am vierten MZP mittlere bis hohe negative Effekte. Die negativen Effekte erklären sich dadurch, dass ein höheres und damit besseres Kompetenzniveau zu einer niedrigeren, aber bessere Klausurnote führt. Damit werden H2.1 und H2.2 bestätigt.

Das zu Studienbeginn erreichte Niveau hat also bereits einen Einfluss auf die erzielten Prüfungsnoten und damit auf den Studienerfolg. Diese

Ergebnisse stellen für die Probanden in den niedrigen Niveaus, die nach Kap. 6.3.1 auch als Risikogruppen bezeichnet werden können, ein großes Problem dar, da es ihnen in der Regel nicht gelingt, die niedrigsten Kompetenzniveaus zu verlassen. In den beiden untersten Kompetenzniveaus (MZP 2) bestehen mehr als die Hälfte der Probanden (Niveau 0: 61% und Niveau I: 58%) die Klausur der Technischen Mechanik am Ende des ersten Fachsemesters (›TM 1‹) nicht. In dem niedrigsten Niveau I am dritten MZP sind es sogar 77%. Ähnliche Ergebnisse ergeben sich für die Klausur am Ende des zweiten Fachsemesters. 57% bzw. 50% der Probanden der niedrigsten Niveaus an MZP 2 bestehen die Klausur ›TM 2‹ nicht. Probanden des Niveaus I an MZP 3 (65%) und 4 (46%) bestehen ebenfalls nicht.

Die in Kap. 6.3.1 genannten Prädiktoren für das Erreichen eines Kompetenzniveaus zu Studienbeginn, haben somit auch eine Aussagekraft für den Studienerfolg. Zu den Prädiktoren gehören die Wahl des Physik-Kurses in der Sekundarstufe II, die Art des Erwerbs der Hochschulzugangsberechtigung, die HZB-Note, der Migrationshintergrund (Muttersprache) und der Bildungshintergrund.

Die gleiche Untersuchung wird nun auch für die erreichten Kompetenzniveaus in den Testinstrumenten der Rechenfähigkeit durchgeführt. Wie u.a. in Kap. 6.3.2 beschrieben, erfassen die Testinstrumente der Rechenfähigkeit ausschließlich mathematische Grundlagen, die bereits in der Schule erlernt werden sollten. Die Inhalte der Mathematikvorlesungen in den ersten beiden Fachsemestern werden dementsprechend nicht berücksichtigt. Eventuelle Korrelationen sind daher so zu interpretieren, dass die mathematischen Grundlagen als Grundvoraussetzung für die Bearbeitung und das Bestehen der Mathematik-Klausur notwendig sind.

Die statistische Überprüfung erfolgt wie bei den Kompetenznvieaus der Technischen Mechanik mit der Spearman-Korrelation, deren Spearman-Korrelationskoeffizienten in Tab. 6.40 zu finden sind.

Auch für die erreichten Kompetenzniveaus der Rechenfähigkeit bestehen signifikante Zusammenhänge mit den Klausurnoten. Bei FUNDAMENT weisen die Niveaus der MZP 2 und 3 signifikante Zusammenhänge mit einem mittleren negativen Effekt mit den Mathematik-Klausurnoten am Ende des ersten Fachsemesters (›Mathematik 1‹)

Tabelle 6.40
Spearman Korrelationskoeffizienten zwischen erreichten Kompetenzniveaus (Rechenfähigkeit) und den Klausurnoten

Niveau	Mathe 1	Mathe 2
FUNDAMENT		
MZP 2	$-.35^{***}$	$-.22$
MZP 3	$-.37^{**}$	$-.09$
MZP 4	$-.59^{***}$	$-.56^{***}$
ALSTER		
MZP 2	$-.48^{***}$	$-.17$
MZP 3	$-.67^{***}$	$-.70^{***}$

Anmerkungen. Mathe 1 = ›Mathematik 1‹; Mathe 2 = ›Mathematik 2‹; MZP = Messzeitpunkt.
** $p < .01$. $^{***}p < .001$

auf. Diese signifikanten Korrelationen bestehen auch in ALSTER, wobei die Effektstärken hier im mittleren und hohen negativen Bereich liegen. Damit wird H2.3 bestätigt.

Für die Klausurnote am Ende des zweiten Fachsemesters (›Mathe 2‹) zeigt sich in FUNDAMENT nur ein signifikanter Zusammenhang (starker negativer Effekt) mit den Kompetenzniveaus am vierten MZP. Zu den Niveaus des dritten MZP gibt es keine signifikante Korrelation. Diese ist jedoch bei ALSTER mit einer hohen negativen Effektstärke vorhanden. Somit kann H2.4 nur eingeschränkt bestätigt werden. Für FUNDAMENT zeigt sich nur ein Zusammenhang zwischen den Niveaus an MZP 4 und der Klausurnote ›Mathe 2‹, während in ALSTER ein Zusammenhang zwischen MZP 3 und der Klausurnote beobachtet werden kann. Daten des vierten MZP liegen für ALSTER nicht vor.

Auch für die Rechenfähigkeit konnten Prädiktoren für das Erreichen eines Kompetenzniveaus zu Studienbeginn ermittelt werden (siehe Kap. 6.3.3). Nach den oben dargestellten Ergebnissen haben diese Prädiktoren auch einen Einfluss auf die Klausurnoten und damit auf den Studienerfolg. Die letzte Fachnote Mathematik und die Note der Hochschulzugangsberechtigung stellen diese Prädiktoren dar.

Abschließend kann die Forschungsfrage 2 mit der Einschränkung bejaht werden, dass am dritten MZP der Studie FUNDAMENT kein

signifikanter Zusammenhang mit der erreichten Klausurnote am Ende des zweiten Fachsemesters besteht.

6.4.3 Forschungsfrage 3

Die dritte Forschungsfrage beschäftigt sich mit der Teilnahme am OSA in der Studienvorphase und deren Auswirkung auf den Studienerfolg (siehe Kap. 3). Wie bereits in Kap. 2.5.1.1 erläutert, wurde in der Studie FUNDAMENT in der Studienvorphase (MZP 1) ein OSA angeboten. Dieses OSA wurde konzipiert, um die Wirksamkeit des OV zu überprüfen (siehe Kap. 2.5.1.1). Dementsprechend wird das OSA zu zwei Zeitpunkten eingesetzt, nämlich vor der Bearbeitung von Teilen oder des gesamten OV - bezeichnet als OSA 1 - und nach der Teilnahme am OV (OSA 2). Der Einfluss des OV wird im Rahmen dieser Arbeit nicht behandelt, entsprechende Ergebnisse finden sich in Pelz et al. (2021). OSA 1 und 2 können in zwei inhaltliche Bereiche (naturwissenschaftliche und mathematische Grundlagen) unterteilt werden (siehe Kap. 2.5.1.1). Die Probandenzahlen liegen zwischen 12 und 33 (siehe Tab. 6.1). Diese Zahlen reduzieren sich weiter, da nicht alle OSA-Teilnehmer auch die entsprechenden Prüfungen am Ende der ersten beiden Fachsemester abgelegt haben. Es soll untersucht werden, ob die bloße Teilnahme am OSA bzw. der dort erzielte Summenscore einen Einfluss auf die Klausurnoten der Technischen Mechanik bzw. Mathematik hat.

Hypothese H3.1 postuliert, dass die Teilnahme am OSA naturwissenschaftliche oder mathematische Grundlagen zu einer signifikant besseren Klausurnote der Technischen Mechanik (›TM 1‹ bzw. ›TM 2‹) mit einem kleinen Effekt im Vergleich zur Kontrollgruppe (Probanden, die nicht an dem entsprechenden OSA teilgenommen haben) beiträgt. Während H3.2 den gleichen Zusammenhang für die Mathematik-Klausurnoten (›Mathe 1‹ bzw. ›Mathe 2‹) postuliert. Die Betrachtung erfolgt getrennt für OSA 1 und OSA 2 jeweils für die beiden Inhaltsbereiche naturwissenschaftliche und mathematische Grundlagen getrennt für die vier Klausuren, sodass eine weitere Verringerung der Stichprobengröße ausgeschlossen werden kann. Es wird auch geprüft, ob das OSA der mathematischen Grundlagen einen Einfluss auf die Technische Mechanik-Klausuren hat. Dies ist auch inhaltlich nachvollziehbar, da

in den Technischen Mechanik-Prüfungen mathematische Kenntnisse vorausgesetzt werden. Der umgekehrte Zusammenhang (Einfluss des OSA naturwissenschaftliche Grundlagen auf die Mathematik-Klausuren) wird ebenfalls untersucht, obwohl es hier keine inhaltlichen Überschneidungen gibt. Ein Mittelwertvergleich für zwei unabhängige Stichproben, wie der t-Test, kann zur Überprüfung von H3.1 und H3.2 verwendet werden. Dabei werden die Mittelwerte der einzelnen Klausurnoten für die Gruppe der OSA-Teilnehmer (unterteilt in OSA 1 oder 2, sowie jeweils naturwissenschaftliche oder mathematische Grundlagen) und derer, die nicht am entsprechenden OSA teilgenommen haben, verglichen. Die Voraussetzungen des t-Tests sind in Kap. 6.3.3 zu finden.

Die erste Bedingung wird durch die zufällige Ziehung der Probanden aus der Grundgesamtheit erfüllt. Innerhalb der Stichprobe beeinflusst ein bestimmter Messwert eines Probanden nicht den Messwert eines anderen Probanden. Auch zwischen den Stichproben beeinflusst der Messwert eines Probanden der einen Stichprobe nicht die Messwerte der anderen Stichprobe. Bereits in Kap. 6.1.2 wurde erwähnt, dass der t-Test robust gegenüber der Verletzung der Normalverteilung ist, sofern die Teilstichproben $n > 30$ sind; dies gilt sowohl für abhängige als auch für unabhängige Stichproben (Bortz & Schuster, 2010; Eid et al., 2017). Wie Tab. 6.41 zeigt, ist n_2 (Teilstichprobengröße mit Absolvierung des OSA) in keinem der Fälle größer als dreißig. Daher wird der Shapiro-Wilk-Test verwendet, um zu prüfen, ob eine Normalverteilung vorliegt. In allen Fällen wird der p-Wert signifikant (bei einem Signifikanzniveau von $\alpha = .05$), sodass die Nullhypothese (es liegt eine Normalverteilung vor) verworfen wird und die entsprechende Bedingung des t-Tests nicht erfüllt ist. Auch die grafische Überprüfung (Histogramm) bestätigt die Ergebnisse des Shapiro-Wilk-Tests. Die Prüfung auf Homoskedastizität (Varianzhomogenität) mit dem Levene-Test ist daher nicht erforderlich. Der t-Test kann nicht verwendet werden und es muss auf einen nichtparametrischen Test ausgewichen werden, der als voraussetzungsärmer gilt (Bortz & Schuster, 2010; Eid et al., 2017; Field et al., 2012).

Das nichtparametrische Äquivalent zum unabhängigen t-Test ist der Mann-Whitney-U-Test (auch Mann-Whitney-Test, Wilcoxon-Mann-Whitney-Test oder Wilcoxon-Rangsummentest genannt) (Bortz & Schus-

ter, 2010; Eid et al., 2017; Field et al., 2012). Beim Mann-Whitney-U-Test werden alle Messwerte in Rangwerte transformiert und zur Bestimmung der zentralen Tendenz nicht die Mittelwerte, sondern die Mediane (Mdn) verwendet (Eid et al., 2017; Field et al., 2012). Die Anwendung des Mann-Whitney-U-Tests setzt voraus, dass das Merkmal in der Population stetig ist und in beiden Teilstichproben die gleiche Verteilungsform vorliegt (Normalverteilung nicht erforderlich) (Eid et al., 2017). Die Stichprobe erfüllt diese Voraussetzungen für alle Fälle. Die Berechnung der Wahrscheinlichkeit (p) kann auf zwei Arten erfolgen: Zum einen gibt es den ›exakten‹ Ansatz, der eine Normalverteilung impliziert und eine Monte-Carlo-Methode zur Berechnung verwendet (Field et al., 2012). Im Gegensatz dazu steht der ›normale‹ Ansatz, der davon ausgeht, dass die Daten nicht normalverteilt sind (Field et al., 2012). Aus zeitökonomischen Gründen empfehlen Field et al. (2012) die Verwendung des ›normalen‹ Ansatzes bei großen Stichprobenumfängen, wobei als Grenze der in der Funktion ›wilcox.test‹ aus dem ›R‹-Paket ›stats‹ (R Core Team, 2023) hinterlegte Wert von $n > 50$ verwendet werden kann. Dieser wird in der Analyse verwendet, da in allen Fällen die Stichproben über dieser Grenze liegen (siehe Tab. 6.41). Field et al. (2012) empfehlen die Verwendung einer Kontinuitätskorrektur zur Korrektur des p-Wertes bei Verwendung des ›normalen‹ Ansatzes, während Eid et al. (2017) die Verwendung empfehlen, wenn die beiden Teilstichprobengrößen stark voneinander abweichen. Die Verwendung der Kontinuitätskorrektur verhindert, dass der p-Wert zu konservativ geschätzt wird (Eid et al., 2017; Field et al., 2012). Da beide Punkte zutreffen, wird die Kontinuitätskorrektur angewendet. Als Effektstärke für den Mann-Whitney-U-Test wird der Pearson Korrelationskoeffizient (r) verwendet (Field et al., 2012). Die Berechnung erfolgt nach der von Field et al. (2012) vorgeschlagenen Rechnung in ›R‹ (Einzelheiten siehe Field et al., 2012, S. 665–666). Die Einteilung der Effektstärke ist in Kap. 6.2.1 zu finden.

Tab. 6.41 zeigt die Ergebnisse des Mann-Whitney-U-Tests für unabhängige Stichproben für den OSA an MZP 1. Zunächst werden die relevanten deskriptiven Statistiken aufgeführt. Dabei wird unterschieden zwischen der Teilstichprobengröße der Probanden, die den OSA

Tabelle 6.41
Ergebnisse des Mann-Whitney-U-Test für unabhängige Stichproben an MZP 1

	n_1	n_2	n	Mdn_1	Mdn_2	U	z	p	r
OSA 1									
NG									
TM 1	247	30	277	5	3.4	4485	−2.32	.020	.14
TM 2	170	21	191	5	2.4	2628	−3.96	<.001	.29
Mathe 1	134	16	150	5	5	1660	−2.48	.013	.15
Mathe 2	56	9	65	5	5	218	−0.24	.807	.02
MG									
TM 1	251	26	277	5	4.5	3770	−1.76	.078	.14
TM 2	175	16	191	5	2.3	2099	−3.73	<.001	.46
Mathe 1	134	16	150	5	5	1264	−1.90	.058	.15
Mathe 2	58	7	65	5	5	185	−0.39	.698	.05

Fortsetzung Tab. 6.41

	n_1	n_2	n	Mdn_1	Mdn_2	U	z	p	r
OSA 2									
NG									
TM 1	263	14	277	5	2.8	2339	−2.15	.032	.13
TM 2	181	10	191	5	2.1	1394	−3.28	.001	.24
Mathe 1	139	11	150	5	3.3	1041	−2.91	.004	.17
Mathe 2	58	7	65	5	5	185	−0.39	.698	.03
MG									
TM 1	265	12	277	5	2.4	2216	−2.74	.006	.22
TM 2	182	9	191	5	2.0	1354	−3.73	<.001	.46
Mathe 1	140	10	150	5	3.3	994	−3.19	.001	.26
Mathe 2	58	7	65	5	5	185	−0.39	.698	.05

Anmerkungen. n_1 = Teilstichprobengröße (ohne Absolvierung OSA); n_2 = Teilstichprobengröße (mit Absolvierung OSA); n = Stichprobengröße ($n = n_1 + n_2$); Mdn_1 = Median von n_1; Mdn_2 = Median von n_2; U = Mann-Whitney-Test-Statistik; z = Z-Statsitik; p = Wahrscheinlichkeit; r = Pearson Korrelationskoeffizient (Effektstärke); OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; TM 1 = ›Technische Mechanik 1‹; TM 2 = ›Technische Mechanik 2‹; Mathe 1 = ›Mathematik 1‹; Mathe 2 = ›Mathematik 2‹.

nicht (n_1) oder erfolgreich (n_2) abgeschlossen haben. Hierbei werden nur die Probanden des Gesamtdatensatzes berücksichtigt, die an der Studie FUNDAMENT teilgenommen haben und für die eine entsprechende Prüfungsnote vorliegt. Die gleiche Unterscheidung gilt für die Mediane (Mdn). Außerdem wird die Stichprobengröße ($N = n_1 + n_2$) angegeben. Anschließend sind die Teststatistiken des Mann-Whitney-U-Tests dargestellt. Dies sind die Mann-Whitney-Test-Statistik (U), die Z-Statsitik (z), die Wahrscheinlichkeit (p) und schließlich der Pearson Korrelationskoeffizient (r). Diese sind entsprechend für OSA 1 bzw. 2 für die Inhaltsbereiche naturwissenschaftliche oder mathematische Grundlagen für die zugehörigen Klausurnoten im ersten (›TM 1‹ und ›Mathe 1‹) bzw. zweiten Fachsemester (›TM 2‹ und ›Mathe 2‹) angegeben. Da nicht alle Probanden, die den OSA absolviert haben, auch die Klausuren im ersten bzw. zweiten Fachsemester absolviert haben, ist die Teilstrichprobengröße n_2 kleiner als die in Tab. 6.1 genannten Stichprobe.

Die Betrachtung der Ergebnisse beginnt zunächst mit dem Einfluss auf die Technische Mechanik-Klausurnoten. Der Mann-Whitney-U-Test zeigt, dass insbesondere für den Inhaltsbereich der naturwissenschaftlichen Grundlagen signifikante Unterschiede in den Technischen Mechanik-Klausurnoten zwischen den Gruppen bestehen. Dies bedeutet, dass Probanden, die an einem OSA für den Inhaltsbereich naturwissenschaftliche Grundlagen teilgenommen haben, in den Technischen Mechanik-Klausuren des ersten bzw. zweiten Fachsemesters bessere Klausurnoten erzielen als Probanden, die an keinem OSA naturwissenschaftliche Grundlagen teilgenommen haben. Die Effekte liegen zwischen $r = .13$ und $.26$ und sind damit als gering einzustufen. Die Teilnahme am OSA 2 MG hat ebenfalls einen kleinen Effekt ($r = .22$) auf die ›TM 1‹-Klausurnote. Nur die Teilnahme am OSA 1 MG hat keinen Effekt auf die ›TM 1‹-Noten. Die Teilnahme am OSA 1 bzw. 2 MG zeigt ebenfalls eine Verbesserung der ›TM 2‹-Note im zweiten Fachsemester gegenüber der Kontrollgruppe, die Effektstärke liegt mit $r = .46$ jeweils im mittleren Bereich.

Im Folgenden wird der Einfluss der Teilnahme am OSA auf die Mathematik-Klausurnoten untersucht. Für den Inhaltsbereich der na-

turwissenschaftlichen Grundlagen zeigt sich, dass die Teilnehmer am OSA 1 oder 2 bessere Mathematik-Klausurnoten im ersten Fachsemester (›Mathe 1‹) erzielen mit einer geringen Effektstärke von $r = .15$ bzw. .17 gegenüber der Kontrollgruppe. Für den Inhaltsbereich mathematische Grundlagen kann mit Ausnahme des OSA 2 bezüglich der Mathematik-Klausur in dem ersten Fachsemester (›Mathe 1‹) kein Einfluss auf die Mathematiknoten der ersten beiden Fachsemester festgestellt werden. Der vorhandene Effekt von OSA 2 MG auf die Klausurnote ›Mathe 1‹ ist als gering einzustufen ($r = .26$).

Insgesamt ist auffällig, dass alle Mdn_1 bei der Note 5 liegen (siehe Tab. 6.41). Es ist zu beachten, dass die Mittelwerte nicht alle bei 5 sind, sondern darunter, und somit einer besseren Note entsprechen.

Die dargestellten Ergebnisse widerlegen die Hypothese H3.1, da gezeigt werden konnte, dass die Teilnahme an einem OSA (naturwissenschaftliche oder mathematische Grundlagen) nicht in allen Fällen zu signifikant besseren Noten in den Klausuren der Technischen Mechanik der ersten beiden Fachsemester führt. Überraschend ist, dass die Teilnahme am OSA 1 MG nicht zu einer signifikant besseren Note in der Klausur ›TM 1‹ führt, aber einen mittleren Effekt auf die Klausurnote der ›TM 2‹ ausübt.

H3.2 muss ebenfalls falsifiziert werden. Es zeigt sich zwar ein Einfluss von OSA 1 NG sowie OSA 2 NG und MG auf die Mathematiknote im ersten Fachsemester, nicht jedoch von OSA 1 MG. Ebenso kann keine Auswirkung des OSA auf die Mathematiknote nach dem zweiten Fachsemester (›Mathe 2‹) festgestellt werden. Erstaunlich ist wiederum, warum der OSA 1 MG erneut keinen Effekt auf die Klausur im ersten Fachsemester, in diesem Fall die Mathematik-Klausur (›Mathe 1‹), hat. Warum kein Einfluss auf die Note der Mathematik-Klausur im zweiten Fachsemester nachgewiesen werden kann, ist ebenfalls unklar.

Die Interpretation dieser Ergebnisse muss auf zwei Ebenen erfolgen, zum einen inhaltlich und zum anderen teststatistisch. Auf der inhaltlichen Ebene steht die Kausalität im Vordergrund. Der OSA wurde konzipiert, um mögliche Auswirkungen der Teilnahme am OV zu ermitteln (Pelz et al., 2021). Es ist jedoch fraglich, ob die bloße Teilnahme am OSA - Teile des OV wurden nur von 5 (NG) bzw. 8 (MG) Probanden

absolviert - zu einer Verbesserung der Leistungen in den relevanten Prüfungen beitragen kann. Als Feedback erhielten die Probanden nicht die Lösungen der bearbeiteten OSA-Aufgaben, sondern eine Handlungs- bzw. Lernempfehlung in Form einer Übersicht der zu bearbeitenden Themenbereiche mit den jeweils erreichten Prozentzahlen (siehe Abb. 2.6). Es ist daher auszuschließen, dass durch bloßes Auswendiglernen einer vorgeschlagenen Lösung eine Verbesserung erzielt werden kann. Es zeigt sich auch, dass die Bearbeitung des OV nicht zwangsläufig zu einer Verbesserung des Summenscores von OSA 1 auf OSA 2 führt (Pelz et al., 2021). Es ergibt sich kein eindeutiges Bild, vielmehr sind minimale Verbesserungen, aber auch Verschlechterungen zu verzeichnen (Pelz et al., 2021). Es ist daher fraglich, ob die signifikanten Ergebnisse tatsächlich kausal auf die Teilnahme am OSA zurückzuführen sind. Dafür spricht auch, dass zwar ein Einfluss der OSA naturwissenschaftlicher und mathematischer Grundlagen auf die Technische Mechanik-Klausuren plausibel erscheint, allerdings die signifikanten Ergebnisse bzgl. OSA naturwissenschaftlicher Grundlagen auf die Note der Mathematik-Klausur im zweiten Fachsemester inhaltlich nicht erklärbar sind. Vielmehr ist davon auszugehen, dass motivationale Effekte zu den Ergebnissen beitragen. Wie bereits in Kap. 2.3 erläutert, nutzen vor allem Studierende derartige Angebote, die diese eigentlich nicht benötigen würden, da sie bereits zu den leistungsstarken Studierenden gehören. Es ist daher davon auszugehen, dass sie auch ohne die Teilnahme an entsprechenden Angeboten bessere Klausurnoten erbringen. Umgekehrt formuliert: Besonders leistungsschwache Studierende könnten sich durch die Nutzung derartiger Angebote verbessern, nehmen diese aber nur in geringer Zahl in Anspruch. Dies könnte zu einer Verzerrung der Ergebnisse führen und die Aussagekraft entsprechend einschränken.

Auf teststatistischer Ebene können die Auswirkungen der geringen Teilnehmerzahl der OSA und die daraus resultierenden Verteilungsverhältnisse diskutiert werden. Grundsätzlich waren höhere Teilnehmerzahlen bei der Datenerhebung erstrebenswert, konnten aber aus verschiedenen Gründen nicht erreicht werden (siehe Kap. 2.5.1.1). Nicht signifikante Ergebnisse lassen sich auch auf eine zu geringe Stichprobengröße zurückzuführen, die zu einer geringen Teststärke führt (Döring &

Bortz, 2016). Daher bietet es sich an, die Teststärke (engl.: *power*) mit Hilfe einer Poweranalyse zu überprüfen. Vereinfacht kann die Teststärke als Sensitivität der Untersuchung verstanden werden, einen vorhandenen Effekt zu erkennen (Bortz & Schuster, 2010). Etwas detaillierter dargestellt ist $1 - \alpha$ die Wahrscheinlichkeit richtigerweise die Nullhypothese (H_0) beizubehalten, also die Wahrscheinlichkeit, überhaupt feststellen zu können, dass kein Effekt vorliegt (Eid et al., 2017). Bezogen auf die Teststärke $(1 - \beta)$ wird die Alternativhypothese (H_1) betrachtet (Eid et al., 2017). Während β die Wahrscheinlichkeit für eine falsche Entscheidung gegen H_1 ist, kann $1 - \beta$ als Wahrscheinlichkeit verstanden werden, sich richtigerweise für H_1 zu entscheiden und damit einen korrekten postulierten Populationseffekt aufzudecken (Eid et al., 2017). Die Teststärke $(1 - \beta)$ hängt von der Stichprobengröße (n), dem Signifikanzniveau (α) und der erwarteten Effektstärke ab (Field et al., 2012). Die Teststärke nimmt zu, wenn sich einer der drei Werte erhöht (Döring & Bortz, 2016). Die Poweranalyse kann *a-priori* oder *post-hoc* durchgeführt werden (Döring & Bortz, 2016; Eid et al., 2017; Field et al., 2012). A-priori-Poweranlysen werden, wie der Name schon sagt, vor der Testdurchführung erstellt und geben die Stichprobengröße in Abhängigkeit vom Signifikanzniveau, der erwarteten Effektstärke und der gewünschten Teststärke an (Döring & Bortz, 2016; Eid et al., 2017; Field et al., 2012). Damit steht bereits vor der Testdurchführung fest, welche Anzahl an Probanden erreicht werden muss, um die gewünschte Teststärke zu erreichen. Die andere Variante ist die Post-hoc-Poweranalyse, die nach der Testdurchführung erfolgt (Döring & Bortz, 2016; Eid et al., 2017; Field et al., 2012). Hierbei wird die Teststärke auf Basis des Signifikanzniveaus, der erreichten Stichprobengröße und der Effektstärke berechnet (Döring & Bortz, 2016; Eid et al., 2017; Field et al., 2012). Auf die Verwendung der berechneten Effektstärke (retrospektive Teststärkenanalyse), in diesem Fall des Mann-Whitney-U-Test (siehe Tab. 6.41), soll hier verzichtet werden, da dadurch die wahre Teststärke verzerrt geschätzt wird (Döring & Bortz, 2016; Eid et al., 2017; Sedlmeier & Renkewitz, 2018). Vielmehr sollte die Effektstärke daher auf Grundlage theoretischer Vorannahmen festgelegt werden (Döring & Bortz, 2016; Eid et al., 2017). Im Rahmen dieser Arbeit

wird auf die in den Hypothesen H3.1 und H3.2 postulierten kleinen Effektstärken ($r = .1$) zurückgegriffen. Zum Vergleich sind in Tab. 6.42 auch die entsprechenden Teststärken für mittlere und große Effekte angegeben; wie bereits erwähnt, steigt die Teststärke definitionsgemäß mit der Größe der Effekte. Das Signifikanzniveau wird wiederum mit $\alpha = .05$ festgelegt.

Tabelle 6.42
Teststärken der Post-hoc-Poweranalyse (Mann-Whitney-U-Test) für MZP 1

	$1-\beta$					
	OSA 1			OSA 2		
	r			r		
	.1	.3	.5	.1	.3	.5
NG						
TM 1	.13	.45	.81	.10	.28	.55
TM 2	.11	.35	.68	.09	.23	.44
Mathe 1	.10	.29	.57	.09	.24	.46
Mathe 2	.08	.20	.38	.08	.18	.33
MG						
TM 1	.12	.41	.76	.09	.26	.50
TM 2	.10	.30	.59	.09	.21	.41
Mathe 1	.10	.29	.58	.09	.23	.44
Mathe 2	.08	.18	.33	.08	.18	.33

Anmerkungen. $1-\beta$ = Teststärke; r = Pearson Korrelationskoeffizient (Effektstärke); OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; TM 1 = ›Technische Mechanik 1‹; TM 2 = ›Technische Mechanik 2‹; Mathe 1 = ›Mathematik 1‹; Mathe 2 = ›Mathematik 2‹.

Die gewünschte Teststärke sollte bei $1 - \beta = 0.80$ bis 0.95 liegen, wird die untere Grenze gewählt, so können auch bei kleinen Stichproben noch Effekte nachgewiesen werden (Eid et al., 2017). Die Poweranalyse wird mit dem Programm ›G*Power‹ durchgeführt (Faul et al., 2007). Die Ergebnisse zeigen, dass für einen kleinen Effekt $r = .1$ die Teststärke zwischen $1 - \beta = .08$ und .13 liegt und damit die gewünschte Teststärke ($1 - \beta = .8$) deutlich verfehlt wird. Auch bei größeren Effekten, die definitionsgemäß zu größeren Teststärken führen, wird nur einmal (OSA 1

NG TM 1 mit einer Teststärke von .81 bei $r = .5$) der gewünschte Teststärkenwert erreicht. Dies verdeutlicht, dass die geringe Probandenzahl der OSA und die daraus resultierenden Verteilungsverhältnisse die Teststärke und damit die Aussagekraft der Analyse schwächen. Die geringe Teststärke führt zu einer geringen Wahrscheinlichkeit, echte Effekte in der Teilstichprobe zu identifizieren bzw. es können falsch-positive Ergebnisse auftreten, da das signifikante Ergebnis keinen echten Effekt widerspiegelt. Für den vorliegenden Fall, dass die Teststärke durchgängig als zu gering einzustufen ist, empfehlen Döring und Bortz (2016), die Untersuchung als ›nicht ausreichend‹ (engl.: *underpowered*) zu klassifizieren. Dies hat zur Folge, dass die Hypothesen H3.1 und H3.2 nicht aussagekräftig beantwortet werden können (Döring & Bortz, 2016).

Auch die Reliabilitäten der einzelnen Testinstrumente stützen die oben getroffenen Aussagen. Cronbach´s α – Interpretation nach Blanz (2015) (siehe Kap. 4.1.2) – ist für OSA 1 NG (Cronbach´s $\alpha = .44$) als inakzeptabel, für OSA 2 NG (Cronbach´s $\alpha = .67$) als fragwürdig einzustufen. Für die mathematischen Grundlagen zeigen sich mit Werten zwischen .71 und .75 bessere Werte, die als akzeptabel interpretiert werden können. Allerdings weist Streiner (2003) darauf hin, dass Skalen mit mehr als 20 Items zu akzeptablen Cronbach´s α Werten tendieren, auch wenn mehrere Dimensionen vorhanden sind.

H3.3 vertieft die Untersuchung der vorangegangenen Hypothesen. Während zuvor untersucht wurde, ob es einen Einfluss der Teilnahme am OSA auf die Klausurnoten gibt, wird nun geprüft, ob die erzielten Ergebnisse (Summenscores) in einem OSA mit dem Studienerfolg (Klausurnoten) korrelieren. Die Produkt-Moment-Korrelation nach Pearson könnte zur Berechnung der Korrelation zwischen den beiden Variablen verwendet werden. Für die Anwendung müssen mehrere Bedingungen erfüllt sein. Die Daten müssen mindestens intervallskaliert sein (Field et al., 2012) und es sollten keine Ausreißer im Datensatz vorhanden sein, da die Korrelation sensitiv auf diese reagiert (Eid et al., 2017). Diese beiden Bedingungen sind im vorliegenden Fall erfüllt. Allerdings muss auch ein linearer Zusammenhang zwischen den beiden Variablen bestehen und sie müssen bivariat normalverteilt sein (Field et al., 2012). Beide Bedingungen sind nicht erfüllt. Durch das häufige Auftreten der Klausurnote 5

ist die Verteilung stark verzerrt, sodass nicht von einem linearen Zusammenhang gesprochen werden kann. Außerdem kann die Annahme einer bivariaten Normalverteilung angezweifelt werden. Die Überprüfung mit dem ›MVN‹-Paket (Korkmaz et al., 2014) in ›R‹ ergibt für verschiedene Verfahren (›Mardia‹, ›Henze-Zirkler‹ und ›Royston‹) widersprüchliche Ergebnisse. Auch die grafische Überprüfung (Histogramm) deutet nicht auf eine Normalverteilung hin. Wie bereits in Kap. 6.1.2 erläutert, könnte bei einer Stichprobengröße von $n > 30$ ohne weitere Überprüfung von einer Normalverteilung ausgegangen werden. Die Stichprobengröße der Probanden, die den OSA absolviert haben (n_2) und somit einen Summenscore aufweisen, ist jedoch in keinem Fall größer als 30 (siehe Tab. 6.41). Die Annahme einer bivariaten Normalverteilung wird daher verworfen. Field et al. (2012) empfehlen in diesem Fall u.a. die Verwendung eines alternativen Korrelationskoeffizienten.

Alternativ zur Produkt-Moment-Korrelation nach Pearson kann die Spearman-Korrelation verwendet werden (siehe Kap. 6.4.2). Es handelt sich um eine nichtparametrische Statistik, die auch dann angewendet werden kann, wenn z.B. die Variablen nicht bivariat normalverteilt sind (Field et al., 2012).

In Tab. 6.43 sind die Ergebnisse der Korrelationsanalyse für MZP 1 dargestellt. Die dargestellten Zahlen sind die Spearman-Korrelationskoeffizienten (r_s) zwischen den Summenscores (TS) im OSA und den Klausurergebnissen. Es gibt nur zwei signifikante Zusammenhänge. Zwischen den erreichten Summenscores in OSA 1 für den Inhaltsbereich naturwissenschaftliche Grundlagen und den Klausurergebnissen in der Technische Mechanik-Klausur im zweiten Fachsemester (TM 2) besteht eine Korrelation mit mittlerer Effektstärke ($r_s = -.39$). Es besteht sogar ein großer Effekt ($r_s = -.51$) zwischen dem in OSA 1 MG erreichten Summenscores und den Klausurergebnissen der TM 1-Klausur.

H3.3 muss falsifiziert werden, da es nur zwei signifikante Korrelationen zwischen den Summenscores im OSA und den Klausurergebnissen gibt. Wie schon bei der Interpretation der Ergebnisse von H3.1 und H3.2 müssen die Gründe dafür analysiert werden. Die genannten signifikanten Korrelationen entsprechen zwar den Erwartungen, es sind jedoch weitere signifikante Zusammenhänge zu erwarten. Die Gründe können

Tabelle 6.43
Spearman Korrelationskoeffizienten an MZP 1 zwischen den Summenscores im OSA und den Klausurnoten

	TM 1	TM 2	Mathe 1	Mathe 2
OSA 1				
NG	−.24	−.39*	−.35	−.56
MG	−.51**	−.32	−.40	−.52
OSA 2				
NG	−.25	−.06	−.49	−.43
MG	−.27	−.20	−.27	−.52

Anmerkungen. TM 1 = ›Technische Mechanik 1‹; TM 2 = ›Technische Mechanik 2‹; Mathe 1 = ›Mathematik 1‹; Mathe 2 = ›Mathematik 2‹; OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen.
* $p < .05$. ** $p < .01$.

wiederum auf inhaltlicher oder teststatistischer Ebene liegen. Die inhaltliche Argumentation erfolgt analog zu den Hypothesen H3.1 und H3.2, deswegen wird an dieser Stelle auf die Wiederholung verzichtet. Eine erneute Post-hoc-Poweranalyse kann teststatisch Aufklärung über die Teststärke in Bezug auf die berechneten Spearman-Korrelationen geben. Die Poweranalyse wird wiederum mit dem Programm ›G*Power‹ (Faul et al., 2007) durchgeführt. Die zuvor verwendeten Parameter werden beibehalten: kleine Effektstärken ($r = .1$) und Signifikanzniveau ($\alpha = .05$). Zur besseren Vergleichbarkeit werden auch wieder die Teststärken für mittlere und große Effekte angegeben (siehe Tab. 6.44).

Die Ergebnisse aus Tab. 6.44 zeigen, dass die Teststärke für kleine Effekte ($r = .1$) zwischen $1 - \beta = .08$ und .13 und damit deutlich unter der gewünschten Mindestteststärke ($1 - \beta = .8$) liegt. Bei den hohen Effekten wird die Mindestteststärke zweimal (mit $1 - \beta = .85$ und .9) überschritten, derart hohe Effekte wären also auch bei der erreichten Stichprobengröße aussagekräftig gewesen. Im Vergleich mit der Post-hoc-Poweranalyse des Mann-Whitney-U-Tests ergeben sich für kleine Effektstärken die gleichen Ergebnisse (siehe Tab. 6.42 und 6.44). Für mittlere und hohe Effektstärken gibt es Unterschiede, die Post-hoc-Poweranalyse der Spearman-Korrelation zeigt häufig etwas höhere

Teststärken. Dennoch sind die Teststärken nicht ausreichend, sodass auch die Untersuchung der Spearman-Korrelation als ›underpowered‹ eingestuft werden muss. Somit kann auch die Hypothese H3.3 nicht aussagekräftig beurteilt werden.

Tabelle 6.44
Teststärken der Post-hoc-Poweranalyse (Spearman-Korrelation) für MZP 1

	$1-\beta$					
	OSA 1			OSA 2		
	r			r		
	.1	.3	.5	.1	.3	.5
NG						
TM 1	.13	.50	.90	.10	.28	.60
TM 2	.11	.38	.78	.08	.22	.46
Mathe 1	.10	.31	.66	.09	.23	.50
Mathe 2	.08	.20	.42	.08	.16	.33
MG						
TM 1	.12	.45	.85	.09	.25	.54
TM 2	.10	.31	.66	.08	.20	.42
Mathe 1	.10	.31	.66	.08	.22	.46
Mathe 2	.08	.16	.33	.08	.16	.33

Anmerkungen. $1-\beta$ = Teststärke; r = Pearson Korrelationskoeffizient (Effektstärke); OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen; TM 1 = ›Technische Mechanik 1‹; TM 2 = ›Technische Mechanik 2‹; Mathe 1 = ›Mathematik 1‹; Mathe 2 = ›Mathematik 2‹.

Aufgrund der geringen Stichprobengröße an MZP 1 kann die Forschungsfrage 3, ob die Teilnahme und die Ergebnisse im OSA (FUNDAMENT) einen Einfluss auf den Studienerfolg haben, nicht abschließend beantwortet werden. Gleiches gilt für die drei postulierten Hypothesen, die zwar zum Teil signifikante Ergebnisse liefern, jedoch aufgrund der geringen Teststärken allesamt als ›underpowered‹ einzustufen sind. Dementsprechend sind die dargestellten Ergebnisse eher als Indizien denn als belastbare Aussagen zu verstehen.

6.4.4 Forschungsfrage 4

Während in der dritten Forschungsfrage die in FUNDAMENT entwickelten digitalen Elemente der Studienvorphase hinsichtlich ihres Einflusses auf die Klausurnoten am Ende der ersten beiden Fachsemester im Bauingenieurwesen untersucht wurden, rückt in Forschungsfrage 4 wieder die Studieneingangsphase in den Fokus. Es soll untersucht werden, ob die für die Studieneingangsphase entwickelten iOM einen Einfluss auf das erreichte Kompetenzniveau in der Technischen Mechanik haben. Da in Kap. 6.4.2 gezeigt werden konnte, dass die Kompetenzniveaus signifikant mit den Klausurnoten korrelieren, wäre dementsprechend bei signifikanten Korrelationen zwischen den Kompetenzniveaus und der iOM-Nutzung, der Einfluss der iOM auf die Klausurnoten nachgewiesen.

Die iOM wurden für das erste und zweite Fachsemester entwickelt, jeweils mit Bezug zur zugehörigen Lehrveranstaltung der Technischen Mechanik (siehe Kap. 2.5.1.2). Zusätzlich zu JACK-Übungsaufgaben stehen den Studierenden Lernvideos zum Selbststudium zur Verfügung. Für alle einzelnen JACK-Übungsaufgaben und Lernvideos wurden in den Moodle-Kursen der Technischen Mechanik Verlinkungen erstellt. Jeder Klick eines Probanden auf einen entsprechenden Link wurde protokolliert, sodass die Aufrufe (Klicks) der iOM vorliegen. Entsprechend wird die Nutzung der iOM über diese Aufrufe operationalisiert. Dabei ist eine getrennte Betrachtung der Gesamtaufrufe der iOM eines Semesters sowie der Aufrufe der JACK-Aufgaben und der Lernvideos möglich.

Signifikante Zusammenhänge werden analog zu Forschungsfrage 2 (siehe Kap. 6.4.2) und 3 (siehe Kap. 6.4.3) mit der Spearman-Korrelation berechnet. Die durch die Korrelation berechneten Spearman-Korrelationskoeffizienten sind in Tab. 6.45 zu finden.

Es bestehen keine signifikanten Zusammenhänge zwischen den erreichten Kompetenzniveaus (Technische Mechanik) an MZP 2 und der iOM-Nutzung im ersten Fachsemester (Gesamtaufrufe, JACK-Übungsaufgaben und Lernvideos). Am dritten MZP bestehen signifikante Korrelationen mit mittlerer Effektstärke, sowohl für die Gesamtaufrufe als auch für die Aufrufe der JACK-Übungsaufgaben im ersten Fachsemester.

Für das zweite Fachsemester bestehen ebenfalls signifikante Korrelationen zwischen dem erreichten Kompetenzniveau an MZP 3 und

Tabelle 6.45
Spearman Korrelationskoeffizienten zwischen erreichten Kompetenzniveaus (Technische Mechanik) und der iOM-Nutzung

Niveau	Gesamt	JACK	Videos
	1. Fachsemester		
MZP 2	.10	.10	−.02
MZP 3	.38***	.38***	.01
	2. Fachsemester		
MZP 3	.26*	.30*	−.08
MZP 4	.36***	.37***	.10

Anmerkungen. iOM = interaktive Online-Module; Gesamt = Gesamtaufrufe iOM; JACK = Aufrufe JACK-Übungsaufgaben; Videos = Aufrufe Lernvideos; MZP = Messzeitpunkt.
*$p < .05$. ***$p < .001$

den Gesamtaufrufen (mittlerer Effekt) sowie den Aufrufen der JACK-Übungsaufgaben (kleiner Effekt). Für den vierten MZP ergeben sich ebenfalls signifikante Korrelationen mit mittlerer Effektstärke zwischen den erreichten Kompetenzniveaus und den Gesamtaufrufen bzw. Aufrufen der JACK-Übungsaufgaben.

Die separate Betrachtung der Nutzung der Lernvideos zeigt an keinem MZP einen signifikanten Einfluss auf das erreichte Kompetenzniveau. Generell ist festzustellen, dass sich die Spearman-Korrelationskoeffizienten der Gesamtaufrufe und der Aufrufe der JACK-Übungsaufgaben nur geringfügig unterscheiden. Demnach haben die Aufrufe der JACK-Übungsaufgaben einen größeren Einfluss auf die Gesamtaufrufe als die Aufrufe der Lernvideos. Dies ist jedoch nicht verwunderlich, da es deutlich mehr JACK-Übungsaufgaben als Lernvideos gibt. So stehen für die Technische Mechanik 1 insgesamt 77 JACK-Übungsaufgaben und 4 Lernvideos zur Verfügung, für die Technische Mechanik 2 sind es 57 JACK-Übungsaufgaben und 7 Lernvideos.

H4.1 muss falsifiziert werden, da die erreichten Kompetenzniveaus an MZP 2 keinen signifikanten Zusammenhang mit der iOM-Nutzung im ersten Fachsemester aufweisen. Dagegen können diese Zusammenhänge für die erreichten Kompetenzniveaus am dritten MZP mit einem mittleren Effekt nachgewiesen werden. Generell gilt jedoch die Einschränkung,

dass die Lernvideos bei keinem MZP einen Einfluss zeigen.

Da die Lernvideos auch im zweiten Fachsemester keinen Einfluss auf die erreichten Kompetenzniveaus an den MZP 3 und 4 zeigen, muss auch H4.2 verworfen werden.

Dagegen scheint die Nutzung der JACK-Übungsaufgaben ein guter Indikator für das erreichte Kompetenzniveau an den MZP 3 und 4 zu sein. Bei einer hohen Nutzung der JACK-Übungsaufgaben erreichen die Probanden hohe Kompetenzniveaus. Es könnte aber auch umgekehrt argumentiert werden: Probanden, die ein hohes Kompetenzniveau erreichen, also als leistungsmäßig bessere Probanden einzustufen sind, nutzen auch häufiger die JACK-Übungsaufgaben. Somit könnte u.a. die Nutzung der iOM dazu geführt haben, dass die Probanden die hohen Kompetenzniveaus erreichen. Es könnte aber auch die in Kap. 2.2.1 gemachte Aussage gestützt werden, dass die Probanden die derartige Unterstützungsmaßnahmen eigentlich nicht bräuchten, sie trotzdem nutzen. Im Umkehrschluss bedeutet dies, dass die Probanden, die den Bedarf hätten, sich also auf einem niedrigeren Kompetenzniveau befinden, derartige Unterstützungsangebote trotz Bedarf nicht nutzen.

Zusammenfassend lässt sich festhalten, dass die Forschungsfrage 4 mit Einschränkungen bejaht werden kann. Zum einen zeigen sich keine signifikanten Zusammenhänge zwischen den erreichten Kompetenzniveaus am zweiten MZP und der Nutzung der iOM, zum anderen haben die Lernvideos keinen signifikanten Effekt auf die Kompetenzniveaus. Dafür kann die Nutzung der JACK-Übungsaufgaben als guter Indikator für das Erreichen eines höheren Kompetenzniveaus an den MZP 3 und 4 angesehen werden.

Kapitel 7

Resümee

Im Rahmen dieser Arbeit wurde ein Kompetenzniveaumodell der Technischen Mechanik für die Studieneingangsphase entwickelt. Ziel ist die möglichst frühzeitige Identifizierung von Studienanfängern, die einer Risikogruppe im Sinne eines möglichen Studienabbruchs angehören.

Da der aktuelle Bedarf an Ingenieuren nicht gedeckt werden kann, müssen die Hochschulen neue Ingenieure ausbilden. In den Ingenieurwissenschaften ist nicht nur die Zahl der Studienanfänger rückläufig, auch die hohe Zahl der Studienabbrecher stellt ein großes Problem dar. Die Hochschulen müssen daher die Studienanfänger unterstützen, um einen möglichen Studienabbruch zu vermeiden. Dies wirft die Frage auf, welche Faktoren zu einem Studienabbruch führen und wann dies in der Regel der Fall ist. Es zeigt sich, dass 42% der Studienabbrecher in den Ingenieurwissenschaften ihr Studium in der Studieneingangsphase (erste beiden Fachsemester) vorzeitig beenden. Die Gründe für dem Studienabbruch liegen zumeist im mangelnden Studienerfolg, der häufig durch Leistungsprobleme vor allem in den Grundlagenfächern wie der Technischen Mechanik oder Mathematik verursacht wird. Als Ursache wird ein Rückgang spezieller fachlicher, auch mathematischer Kenntnisse bei den Studienanfängern angesehen. Dies hat zur Folge, dass die Studienanfänger ihr Studium mit Wissenslücken in den genannten Bereichen beginnen, die sie selbstständig aufarbeiten müssen, was sich in der Studieneingangsphase als besonders schwierig erweist. Das ›Referenzmodell zur Qualitätssicherung an Fakultäten der Ingenieurwissenschaften‹ von Heublein und In der Smitten (2013) greift diese Problematik auf.

Es besagt u.a., dass ein Bündel von Unterstützungsmaßnahmen zu unterschiedlichen Zeitpunkten im Studienverlauf hilfreich sein kann, um den Studienerfolg zu verbessern.

In Anlehnung an das Referenzmodell wurde im Rahmen des vom BMBF geförderten Verbundforschungprojektes FUNDAMENT ein Förderkonzept entwickelt. Dieses umfasst Maßnahmen, die in der Studienvorphase und in der Studieneingangsphase ansetzen. Für die Studienvorphase wurden ein OSA und ein OV entwickelt, die inhaltlich naturwissenschaftliche (mechanische Inhalte der Schulphysik) und mathematische Grundlagen abdecken (MZP 1). Für die Studieneingangsphase wurden iOM in Form von digitalen JACK-Übungsaufgaben und Lernvideos für die Veranstaltungen Technische Mechanik 1 und 2 in den ersten beiden Fachsemestern erarbeitet. Darüber hinaus wurden in FUNDAMENT Paper & Pencil-Testungen zu Beginn des ersten Fachsemesters (MZP 2) sowie am Ende des ersten (MZP 3) und zweiten Fachsemesters (MZP 4) an der UDE durchgeführt. In diesen Erhebungen wurden Testinstrumente der Technischen Mechanik und der Rechenfähigkeit eingesetzt, die wiederum von der Studie ALSTER adaptiert wurden. In ALSTER wurden diese Testinstrumente zur Technischen Mechanik (Fachwissen und Modellierungsfähigkeit) sowie zur Rechenfähigkeit (mathematische Grundkenntnisse) entwickelt und zur Datenerhebung an der UDE und der RUB eingesetzt.

Die Probandenakquise für den ersten MZP (FUNDAMENT) erwies sich als problematisch. Da sich der MZP in der Studienvorphase befindet, wurden die Testinstrumente des OSA online zur Verfügung gestellt. Das Untersuchungsdesign sieht vor, dass die Probanden einen OSA 1 absolvieren, anschließend Wissenslücken im OV aufarbeiten und abschließend dessen Wirksamkeit an einem OSA 2 überprüfen. Die Studienanfänger sind in der Studienvorphase noch nicht persönlich an der UDE anwesend und konnten daher nur per E-Mail vom Einschreibewesen angeschrieben werden. Von 441 per E-Mail angeschriebenen Studienanfängern haben nur 14% am OSA teilgenommen. Darüber hinaus haben diese die Testinstrumente zumeist nur unvollständig bearbeitet. Auch eine Probandenvergütung i.H.v. 30 EUR stellte keinen ausreichenden Anreiz dar. So haben am OSA 1 31 Probanden teilgenommen, während es im

nachgeschalteten OSA 2 nur noch 14 Probanden waren. Entsprechend wenige Probanden nutzten den OV zur Vorbereitung auf ihr Studium. Aufgrund der geringen Probandenzahl ist eine weitere Datenerhebung in diesem Kontext unumgänglich, um aussagekräftige Ergebnisse formulieren zu können. Ein großes Problem stellt die Erreichbarkeit der Probanden in den Studienvorphase dar. Die vom Einschreibewesen versandten E-Mails wurden an die nach der Einschreibung zugeteilten E-Mail-Adressen der Hochschule geschickt. Es ist zu erwarten, dass einige Studienanfänger diese Adresse vor Studienbeginn noch nicht final eingerichtet haben oder nicht regelmäßig auf neue E-Mails prüfen. Daher muss ein anderer Weg gefunden werden, um die Studienanfänger zu erreichen und zur Teilnahme zu bewegen.

Die anderen drei MZP sollten ursprünglich ebenfalls online durchgeführt werden, da jedoch in der Pilotierungsphase kaum Probanden gewonnen werden konnten, wurde in der Haupterhebung ein Paper & Pencil-Test durchgeführt. Um eine größere Stichprobe zu generieren, wurde zusätzlich zur Haupterhebung ein Jahr später eine weitere Kohorte (Nacherhebung) untersucht. Aufgrund der COVID-19-Pandemie wurde der letzte MZP (Nacherhebung) online mittels LimeSurvey-Befragung durchgeführt. Auch für die MZP in der Studieneingangsphase wurden Probandenvergütungen für die Teilnahme ausgelobt (MZP 2 und 3: je 20 EUR, MZP 4: 30 EUR). Zusätzlich wurde zur Steigerung der Motivation unter den 10 besten Probanden eines MZP ein Semesterbeitrag verlost. Allerdings erwies sich auch hier die Probandenvergütung nicht als ausreichender Anreiz, da insbesondere im Längsschnitt ein hoher Schwund zu beobachten ist. In der Nacherhebung wurde ein weiterer Anreiz durch mögliche Bonuspunkte für die Klausuren der Technischen Mechanik geschaffen. Diese wurden tendenziell gut angenommen, führten aber dennoch – auch aufgrund der kleineren Grundgesamtheit – zu einer geringeren Teilnehmerzahl, als bei der Haupterhebung.

Somit zeigt sich für alle MZP, dass die Probandenakquise ein großes Problem für derartigen Studien darstellt. In weiteren Untersuchungen müsste zunächst geklärt werden, welche Anreize die Studierenden zur Teilnahme bewegen würden. Dies könnte durch strukturierte mündliche (Interviews) oder schriftliche (Fragebögen) Befragungen geschehen.

Die Beschreibung der vorliegenden Stichprobe erfolgt anhand der erhobenen demographischen Variablen, die für 546 Probanden vorliegen. Die Ausschöpfungsquote dieser Angaben liegt bei 57% und 62% für FUNDAMENT, während sie für ALSTER nur 27% beträgt. Die Gesamtstichprobe zeigt eine für das Ruhrgebiet typische heterogene Studierendenschaft. Etwa die Hälfte der Probanden sind Nicht-Muttersprachler, von denen wiederum ein Drittel in einer Selbsteinschätzung angibt, nur gut oder sogar schlechter Deutsch zu sprechen. 80% der Nicht-Muttersprachler sprechen in der Familie überwiegend eine andere Sprache, während 80% im Freundeskreis überwiegend Deutsch sprechen. Während in den Ingenieurwissenschaften 44% der Studierenden eine niedrige oder mittlere Bildungsherkunft haben, sind es in der Stichprobe mit 60% deutlich mehr. Diese Faktoren können sprachliche Probleme im Studium bedingen.

Der Frauenanteil, der in den Ingenieurwissenschaften typischerweise zwischen 12% und 19% (Klöpping et al., 2017) liegt, wird mit 34% deutlich übertroffen. Die Durchschnittsnote der Hochschulzugangsberechtigung liegt bei 2.7. Die Kurswahl in der Sekundarstufe II ist für das Pflichtfach Mathematik ausgeglichen, dagegen belegen 34% der Probanden einen Leistungskurs in Physik, während 56% gar keinen Physik-Kurs belegen.

Signifikante Unterschiede zwischen den beiden Studien gibt es bei den folgenden demographischen Variablen:

- Durchschnittsalter - 20.5 Jahre
  ALSTER 0.7 Jahre jünger als FUNDAMENT
- Deutsch als Muttersprache - 52%
  ALSTER 58% - FUNDAMENT 48%
- Sehr gute Deutschkenntnisse (Nicht-Muttersprachler) - 68%
  ALSTER 77% - FUNDAMENT 63%
- Hochschulzugangsberechtigung
  - Gymnasium - 58%
    ALSTER 65% - FUNDAMENT 53%

  - Gesamtschule - 28%
    ALSTER 26% - FUNDAMENT 30%
  - Ausland - 10%
    ALSTER 5% - FUNDAMENT 12%

- Vorkurs vor dem Studium absolviert - 44%
  ALSTER 54% - FUNDAMENT 38%

In der Stichprobenbeschreibung zeigt sich also das für das Ruhrgebiet typische Bild einer heterogenen Studierendenschaft. Es zeigen sich aber auch signifikante Unterschiede zwischen den beiden Studien. Diese lassen sich als typische Jahrgangs- bzw. Kohortenunterschiede beschreiben, da die Erhebungen zwei Jahre auseinander liegen. Auffällig ist auch der höhere Anteil ausländischer Studierender in der Studie FUNDAMENT. Diese Probanden könnten dazu beitragen, dass im Vergleich zu ALSTER mehr Nicht-Muttersprachler mit weniger guten Deutschkenntnissen vertreten sind.

Nach der Stichprobenbeschreibung wurde der Datensatz aufbereitet und auf Ausreißer und Extremwerte untersucht. Hierzu wird ein visuelles Verfahren mit sogenannten Box-Whisker-Diagrammen verwendet. Bei der Einzelbetrachtung der MZP ergeben sich für die verschiedenen Testinstrumente teilweise Ausreißer nach unten und nach oben. Ausreißer werden aus dem Datensatz entfernt, wenn ihnen mangelnde Sorgfalt bei der Bearbeitung der Testinstrumente unterstellt werden kann. Dieses Kriterium trifft auf keinen der visuell identifizierten Ausreißer zu, sodass alle Probanden im Datensatz verbleiben.

Anschließend wurden die fehlenden Werte im Datensatz genauer untersucht. Die Überprüfung des Datensatzes ergibt, dass zwischen 1% und 7% fehlende Werte an den einzelnen MZP vorliegen. Diese fehlenden Werte entsprechen dem Typ Missing completely at random, wie der Test von Little (1988) zeigt. Diese fehlenden Werte werden mit Hilfe des *Predictive Mean Matching*-Verfahrens (zufallsbedingter Imputationsalgorithmus) imputiert, sodass keine fehlenden Werte mehr im Datensatz vorhanden sind. Hat ein Proband jedoch an einem MZP nicht teilgenommen, so werden die Daten für diesen MZP nicht imputiert. Der imputierte Datensatz wird mit dem ursprünglichen Datensatz

verglichen, und zwar mit einem t-Test, der die Mittelwertunterschiede der Summenscores vergleicht. Der t-Test zeigt signifikante Unterschiede für MZP und Testinstrumente mit einer hohen Anzahl fehlender Werte. Dies ist jedoch nicht verwunderlich, da bei der Imputation die Werte ›0‹ bzw. ›1‹ verwendet werden und somit eine Mittelwertsverschiebung in positive Richtung verfahrensbedingt ist. Die Mittelwertsunterschiede (Imputationsdatensatz - Quelldatensatz) liegen zwischen 0.1 und 0.7 Punkten, wobei der zweite MZP von ALSTER mit 2.6 eine deutliche Veränderung aufweist. Dieser hohe Wert ist die Folge von 1088 fehlenden Werten (7%) im Datensatz. Der imputierte Datensatz wurde nochmals mit Hilfe von Box-Whisker-Diagrammen auf Ausreißer/Extremwerte untersucht. Wiederum werden auffällige Werte gefunden, die jedoch aus den gleichen Gründen im Datensatz verbleiben.

Nach der Datenaufbereitung wurde der Datensatz mit Hilfe der IRT skaliert. Dazu muss zunächst geprüft werden, ob die Skalierung als Gesamtdatensatz zulässig ist. Entsprechend muss untersucht werden, ob sich die Haupt- und die Nacherhebung von FUNDAMENT sowie die Studien FUNDAMENT und ALSTER hinsichtlich der Reihenfolge der Itemschwierigkeiten (1pl-Rasch-Modell) signifikant unterscheiden. Der Vergleich von Haupt- und Nacherhebung der Studie FUNDAMENT zeigt hohe Korrelationen, lediglich der vierte MZP der Nacherhebung weist keine signifikante Korrelation auf. Aus diesem Grund werden die Daten der Nacherhebung für das Testinstrument der Rechenfähigkeit am vierten MZP aus dem Datensatz entfernt. Inhaltlich lässt sich dieser signifikante Unterschied darauf zurückführen, dass in der Nacherhebung kein Paper & Pencil-Test, sondern eine online LimeSurvey-Befragung eingesetzt wurde. Warum diese Probleme jedoch nur bei dem Testinstrument der Mathematik und nicht bei dem Testinstrument der Technischen Mechanik auftreten, kann nicht abschließend geklärt werden. Es kann nur vermutet werden, da die Items der Rechenfähigkeit am Ende des Tests stehen, dass es sich um einen Motivationsverlust handelt. Der Vergleich der Studien FUNDAMENT und ALSTER zeigt für die Testinstrumente der Technischen Mechanik signifikante Zusammenhänge. Diese lassen sich jedoch nicht für die Testinstrumente der Rechenfähigkeit nachweisen. Somit unterscheiden sich die beiden Studien und eine

gemeinsame Skalierung ist nicht möglich. Die inhaltliche Interpretation liegt darin, dass bei FUNDAMENT ein geschlossenes Antwortformat gewählt wurde, während bei ALSTER Items mit offenem Antwortformat verwendet wurden. Der Unterschied im Antwortformat führt somit zu Veränderungen in der Itemschwierigkeit, die sich in unterschiedlichen Rangplätzen widerspiegeln. Die gemeinsame Skalierung der Studie FUNDAMENT für die Testinstrumente der Technischen Mechanik und der Rechenfähigkeit (außer Nacherhebung MZP 4) ist somit zulässig. Gleiches gilt für die beiden Studien zum Testinstrument der Technischen Mechanik, die Testinstrumente der Rechenfähigkeit müssen getrennt skaliert werden.

Nachdem die Skalierbarkeit des Datensatzes geklärt ist, können die Daten IRT skaliert werden. Dazu werden für jedes Testinstrument (Technischen Mechanik und Mathematik) und jeden MZP Modellvergleiche durchgeführt, indem verschiedene IRT-Modelle (1pl, 2pl, 3pl, 1- oder 2-/3-dimensional) skaliert und anschließend verglichen werden. Die MZP werden untereinander mit Ankeritems verankert, sodass die berechneten Parameter auf der gleichen Skala liegen. Eine Verankerung der Testinstrumente der Technischen Mechanik und der Mathematik wäre durch die Berücksichtigung zusätzlicher Dimensionen möglich, wird aber im Rahmen dieser Arbeit nicht vorgenommen. Für alle MZP des Testinstruments der Technischen Mechanik zeigt das 1-dimensionale 2pl-Birnbaum-Modell die beste Passung für die vorliegenden Daten. Gleiches gilt für die Testinstrumente der Rechenfähigkeit für die in FUNDAMENT erhobenen Daten. Lediglich für die Testinstrumente der Rechenfähigkeit der Studie ALSTER liefert das 1-dimensionale Rasch-Modell die beste Passung.

Mit IRT skalierten Daten können Kompetenzniveaumodelle entwickelt werden, wie sie z.B. im schulischen Kontext in Large-Scale-Assessments wie PISA verwendet werden. Hierfür erweist sich die Skalierung der Daten nach dem 2pl-Birnbaum-Modell als ungeeignet, da die Niveaumodellierung die spezifische Objektivität des 1pl-Rasch-Modells erfordert. Daher wird einem Lösungsansatz von (Wilson, 2003) gefolgt, der die Verwendung von Itemschwierigkeiten mit einer Lösungswahrscheinlichkeit von 80% ($P(u_{ij} = 1) = 80\%$) empfiehlt. Die inhaltliche

Interpretation der Niveaus erfolgt anhand der Items, die den jeweiligen Niveaus zugeordnet sind. Neben den Inhalten können keine schwierigkeitsbestimmenden Merkmale identifiziert werden. Hierfür ist das Forschungsdesign als Large-Scale-Assessment nicht geeignet, weshalb weiterführende Untersuchungen z.B. mit wissenschaftlichen Beobachtungsverfahren sinnvoll sein könnten.

Die berechneten Itemschwierigkeiten (bei $P(u_{ij} = 1) = 80\%$) der drei MZP der Technischen Mechanik unterscheiden sich nicht signifikant voneinander. Signifikante Unterschiede der Personenfähigkeiten liegen jedoch zwischen MZP 2 und 3 bzw. MZP 2 und 4 vor. Zwischen MZP 3 und 4 gibt es keinen signifikanten Unterschied in den Personenfähigkeiten. Die bei der Niveaumodellierung sehr hoch angesetzte Lösungswahrscheinlichkeit von 80% hat zur Folge, dass viele Probanden in das niedrigste Niveau eingeordnet werden. Das niedrigste Niveau erstreckt sich im Gegensatz zu den anderen Niveaus nicht über einen Logit, sondern über einen größeren Bereich. Dies liegt daran, dass es keine Items mit einer so geringen Itemschwierigkeit gibt, um weitere Niveaugrenzen zu definieren. Diese Items wären notwendig, um eine inhaltliche Beschreibung der Niveaus vornehmen zu können. Dadurch ist die Differenzierung im niedrigsten Kompetenzniveau nicht so ausgeprägt wie in den höheren Niveaus.

Eine Besonderheit stellt das Testinstrument der Technischen Mechanik an MZP 2 dar. Da die Probanden noch keine Veranstaltung der Technischen Mechanik besucht haben, können sie ihr Wissen nur durch das Fach Physik in der Schule, durch Vorkurse oder im privaten Umfeld erworben haben. Da dieser MZP am Anfang des Studiums liegt, wurde untersucht, ob es Prädiktoren gibt, die einen signifikanten Einfluss auf das erreichte Kompetenzniveau haben. Folgende Prädiktoren konnten für das Testinstrument der Technischen Mechanik an MZP 2 identifiziert werden:

- Belegung Physik-Kurs in der Sekundarstufe II
  Keine Belegung < Grundkurs < Leistungskurs
- Art des Erwerbs der Hochschulzugangsberechtigung
  Sonstige < Gesamtschule < Gymnasium

- Note der Hochschulzugangsberechtigung
  Schlechte HZB-Noten < gute HZB-Noten
- Muttersprache
  Nicht-Muttersprachler < Muttersprachler
- Bildungsherkunft
  Niedrige Bildungsherkunft < hohe Bildungsherkunft

Probanden mit einem Merkmal auf der linken Seite (z.B. kein Physik-Kurs in der Sekundarstufe II) erreichen tendenziell niedrigere Kompetenzniveaus, während Probanden mit einem Merkmalen auf der rechten Seite (z.B. Leistungskurs Physik) höhere Kompetenzniveaus erreichen. Es wurde bereits erwähnt, dass es von MZP 2 zu den anderen beiden MZP einen signifikanten Anstieg der Personenfähigkeit gibt. Dieser Anstieg ist auch bei der Betrachtung des Niveaumodells erkennbar (die Verteilung der Probanden verschiebt sich). Inhaltlich lässt er sich damit begründen, dass die Probanden an MZP 2 ihre Kenntnisse der Technischen Mechanik hauptsächlich aus der Schule mitbringen und erst im ersten Fachsemester ihre erste Veranstaltung zur Technischen Mechanik besuchen. Dementsprechend ist diese Veränderung zu erklären. Der Umstand, dass es keinen signifikanten Unterschied zwischen MZP 3 und 4 gibt, kann hingegen damit erklärt werden, dass sich die Probanden auch im zweiten Fachsemester Wissen aneignen, aber nur soviel, um ihr Kompetenzniveau zu halten. Der Wissenserwerb reicht nicht aus, um den Sprung auf das nächsthöhere Kompetenzniveau zu schaffen. Weiterhin ist zu beobachten, dass einige Probanden Items, die sie bereits an einem vorherigen MZP gelöst haben, am MZP 4 nicht mehr richtig lösen. Dies deutet auf einen ›Vergessenseffekt‹ oder mangelnde Motivation hin. Zusammenfassend kann für die Testinstrumente der Technischen Mechanik festgehalten werden, dass es den Probanden in den niedrigen Niveaus in der Regel nicht gelingt, über die MZP, d.h. im Verlauf ihres Studiums, höhere Kompetenzniveaus zu erreichen. Dagegen sind die Probanden der mittleren Niveaus in der Lage, höhere Niveaus zu erreichen.

Die Kompetenzniveaumodelle der Rechenfähigkeit müssen für die beiden Studien getrennt entwickelt werden. Dabei ist zu beachten, dass

es keine gemeinsame Skala gibt. Daher ist ein direkter Vergleich der beiden Studien hinsichtlich der Rechenfähigkeit nicht möglich. Um dennoch tendenzielle Zusammenhänge benennen zu können, wird das Niveaumodell von ALSTER trotz Skalierung nach dem 1pl-Rasch-Modell ebenfalls mit einer Lösungswahrscheinlichkeit von 80% ($P(u_{ij} = 1) = 80\%$) modelliert.

Die Testinstrumente der Rechenfähigkeit der Studie FUNDAMENT zeigen ähnliche Ergebnisse wie die Testinstrumente der Technischen Mechanik. Die Itemschwierigkeiten zeigen keine signifikanten Unterschiede, während bei der Personenfähigkeit wiederum ein signifikanter Anstieg von MZP 2 nach MZP 3 zu beobachten ist. Die Verschiebung der Verteilung innerhalb des Niveaumodells ist ebenfalls gegeben. Eine inhaltliche Interpretation in Analogie zu den Testinstrumenten der Technischen Mechanik ist jedoch nicht möglich. Schließlich werden in den Testinstrumenten der Rechenfähigkeit grundlegende mathematische Inhaltsbereiche abgefragt, die bereits in der Schule erworben werden sollten. Die abgefragten Inhalte werden in der Regel in den Veranstaltungen der Mathematik in der Studieneingangsphase nicht behandelt oder wiederholt. Die Probanden müssen sich das Wissen also selbstständig oder durch zusätzliche Lernangebote angeeignet haben. Die Tatsache, dass es keinen Anstieg von MZP 3 nach 4 gibt, ist damit zu begründen, dass sich die Probanden zwar neues Wissen aneignen können, aber vermutlich nur ihren Wissensstand aufrechterhalten, um das Niveau zu halten.

Für die Studie ALSTER können aufgrund von Kodierungsproblemen keine Berechnungen für den vierten MZP der Testinstrumente der Rechenfähigkeit durchgeführt werden. Der Vergleich der MZP 2 und 3 zeigt jedoch, dass es keine signifikanten Unterschiede zwischen den Itemschwierigkeiten und den Personenfähigkeiten gibt. Bei der Niveaumodellierung fällt auf, dass mehr als 20% der Probanden den Sprung von dem mittleren zum höchsten Niveau geschafft haben.

Es lassen sich Tendenzen zwischen den beiden Studien hinsichtlich der Testinstrumente der Rechenfähigkeit formulieren, sodass Prädiktoren für den zweiten MZP definiert werden können. So erweist sich erneut die Wahl des Kurses in der Sekundarstufe II als signifikant für das

Erreichen eines Kompetenzniveaus:

- Belegung Mathematik-Kurs in der Sekundarstufe II (Pflichtfach)
  Grundkurs < Leistungskurs
- Note der Hochschulzugangsberechtigung
  Schlechte HZB-Noten < gute HZB-Noten
- Letzte Fachnote Mathematik (FUNDAMENT)
  Schlechte Noten < gute Noten

Diese Aufzählung folgt der gleichen Interpretation wie die vorangegangene Aufzählung für die Technische Mechanik. Die letzte Fachnote Mathematik gilt nur für FUNDAMENT als Prädiktor, da für ALSTER keine entsprechenden Noten vorliegen.

In beiden Studien gelingt es Probanden mit mittlerem Niveau, an einem späteren MZP ein höheres oder sogar das höchste Niveau zu erreichen. Im Gegensatz dazu gelingt es Probanden auf den niedrigsten Niveaus in der Regel nicht, ein höheres Niveau zu erreichen.

Mit Hilfe des Datensatzes und der entwickelten Niveaumodelle können die folgenden vier Forschungsfragen beantwortet werden:

F1 Können die Befunde der Studie ALSTER hinsichtlich der Inhaltsbereiche Technische Mechanik und Rechenfähigkeit in der Studie FUNDAMENT repliziert werden?

F2 Gibt es signifikante Zusammenhänge zwischen den modellierten Kompetenzniveaus (Technische Mechanik bzw. Rechenfähigkeit) und dem Studienerfolg, sodass die Niveaus als Prädiktoren für den Studienerfolg angesehen werden können?

F3 Haben die Teilnahme und die entsprechenden Ergebnisse am OSA an MZP 1 (FUNDAMENT) einen Einfluss auf den Studienerfolg?

F4 Besteht ein Zusammenhang zwischen den modellierten Kompetenzniveaus (Technische Mechanik) an den MZP 2-4 und der Nutzung der iOM im ersten bzw. zweiten Fachsemester?

Während sich die vierte Forschungsfrage ausschließlich auf das Testinstrument der Technischen Mechanik bezieht, beziehen sich die anderen Forschungsfragen auch auf das Testinstrument der Rechenfähigkeit. Forschungsfrage 3 befasst sich mit der Studienvorphase des Bauingenieurwesens (MZP 1), die anderen drei erstrecken sich über die ersten beiden Fachsemester (MZP 2 bis 4).

F1 muss verneint werden, die Befunde von ALSTER unterscheiden sich signifikant von FUNDAMENT. Die Personenfähigkeiten (Technische Mechanik) der einzelnen MZP wurden mit einem t-Test untersucht. Für alle MZP ergeben sich signifikante Unterschiede. Auffallend ist, dass die verglichenen Mittelwerte an den MZP 3 und 4 in FUNDAMENT höher sind als in ALSTER, während es am zweiten MZP umgekehrt ist. Der Grund dafür könnte sein, dass bei der Imputation an MZP 2 für ALSTER sehr viele fehlende Werte imputiert wurden. Bereits bei dem Vergleich des Quelldatensatzes mit dem imputierten Datensatz ergibt sich eine Mittelwertdifferenz von 2.6, während bei FUNDAMENT nur eine Erhöhung von 0.5 vorliegt. Dieser Unterschied könnte der Grund für den höheren Mittelwert der Personenfähigkeit in ALSTER sein. Die Ergebnisse der Testinstrumente der Rechenfähigkeit konnten ebenfalls nicht repliziert werden, da bereits bei der Prüfung der Skalierbarkeit des Datensatzes festgestellt wurde, dass die unterschiedlichen Antwortformate der Items zu signifikanten Unterschieden in der Rangfolge der Itemschwierigkeiten führen.

Mit der Spearman-Korrelation wurden die für die F2 relevanten Zusammenhänge zwischen den Kompetenzniveaus und den Klausurnoten überprüft. Für beide Testinstrumente und Studien zeigen sich signifikante Zusammenhänge. So korrelieren die Kompetenzniveaus der MZP 2 und 3 mit den Klausurnoten am Ende des ersten Fachsemesters. Ebenso gibt es Korrelationen zwischen den Niveaus an den MZP 3 und 4 und den Klausurnoten am Ende des zweiten Fachsemesters.

Da in der zweiten Forschungsfrage bestätigt werden konnte, dass ein signifikanter Zusammenhang zwischen den Kompetenzniveaus der Technischen Mechanik und der Rechenfähigkeit mit den entsprechenden Klausurnoten besteht, können sowohl die erreichten Kompetenzniveaus als auch die Prädiktoren für die Niveaus an MZP 2 zur Identifizierung

möglicher Risikogruppen hinsichtlich des Studienabbruchs herangezogen werden. Da es den Probanden eines niedrigen Niveaus in der Regel nicht gelingt, in ein höheres Niveau aufzusteigen, deutet ein solches Niveau darauf hin, dass die Probanden gefährdet sind, die Prüfungen in den ersten beiden Fachsemestern nicht zu bestehen. Sie gehören damit zu einer Risikogruppe hinsichtlich eines möglichen Studienabbruchs.

Die dritte und vierte Forschungsfrage beziehen sich auf die in FUNDAMENT entwickelten Online-Elemente. In F3 wird untersucht, ob ein signifikanter Zusammenhang zwischen der Teilnahme bzw. den erreichten Summenscores im OSA und den jeweiligen Klausurnoten besteht. Dabei werden die naturwissenschaftlichen Grundlagen im Hinblick auf die Klausuren der Technischen Mechanik und die mathematischen Grundlagen hinsichtlich der Mathematik-Klausuren untersucht. Da kleine Teilstichproben von $n < 30$ vorliegen, wurde das nichtparametrische Äquivalent zum t-Test, der Mann-Whitney-U-Test, verwendet, um Unterschiede zwischen den Probanden, die am OSA teilgenommen haben, und der Kontrollgruppe hinsichtlich der relevanten Klausurnoten zu untersuchen. Es zeigen sich zum Teil signifikante Zusammenhänge zwischen der Teilnahme am OSA und den Klausurnoten, allerdings nicht für alle relevanten Teilbereiche des OSA und auch nicht für alle Prüfungen. Diese zum Teil widersprüchlichen Ergebnisse führten zu Zweifeln an der Teststärke. Aus diesem Grund wurde eine Post-hoc-Poweranalyse durchgeführt, die zu dem Ergebnis kam, dass die Untersuchung als ›nicht ausreichend‹ einzustufen ist. Somit können keine aussagekräftigen Antworten auf den Zusammenhang zwischen der Teilnahme am OSA und den Klausurnoten gemacht werden.

Der zweite Teil von F3, ob es einen signifikanten Zusammenhang zwischen den Ergebnissen im OSA und den Klausurnoten gibt, wurde mit der Spearman-Korrelation berechnet. Diese zeigt nur zwei signifikante Korrelationen, sodass die Teststärke erneut in Frage gestellt werden kann. Daher wurde eine weitere Post-hoc-Poweranalyse durchgeführt, diesmal für die Spearman-Korrelation. Auch diese kommt zu dem gleichen Ergebnis, die Untersuchung muss als ›nicht ausreichend‹ bewertet werden. Somit kann die F3 nicht abschließend beantwortet werden, da die Teststärke aufgrund der geringen Stichprobengröße zu gering ist.

Die in der Studieneingangsphase eingesetzten iOM stehen in F4 im Mittelpunkt. Es wird untersucht, ob es einen signifikanten Zusammenhang zwischen der Nutzung der iOM und den erreichten Kompetenzniveaus in den Testinstrumenten der Technischen Mechanik gibt. Die Operationalisierung der iOM erfolgt über die Klicks (Aufrufe) auf die entsprechenden Elemente im Moodle-Kurs. Dabei wird zwischen den Gesamtaufrufen der iOM, den Aufrufen der JACK-Übungsaufgaben und den Aufrufen der Lernvideos unterschieden. Signifikante Zusammenhänge werden mit der Spearman-Korrelation ermittelt. Die Lernvideos zeigen bei keinem MZP einen signifikanten Zusammenhang zu den erreichten Kompetenzniveaus. Dagegen weisen sowohl die Gesamtaufrufe, wie auch die Aufrufe der JACK-Übungsaufgaben einen signifikanten Zusammenhang zu den Kompetenzniveaus an MZP 3 (iOM: 1. und 2. Fachsemester) und MZP 4 (iOM: 2. Fachsemester) auf. Somit scheint die Nutzung der iOM ein guter Indikator für die erreichten Kompetenzniveaus an den MZP 3 und 4 zu sein.

Insgesamt kann diese Dissertation dazu beitragen, das Verständnis und die Herausforderungen der Studieneingangsphase zu erweitern. Durch die Entwicklung der Kompetenzniveaumodelle können Prädiktoren benannt werden, die dazu beitragen, Studienanfänger mit einem erhöhten Studienabbruchrisiko zu identifizieren. Dies ermöglicht eine frühzeitige Intervention, sodass gezielte Fördermaßnahmen eingesetzt werden können, um den Studienerfolg zu verbessern und Studienabbrüche zu vermeiden. Die Erkenntnisse dieser Dissertation leisten somit einen wichtigen Beitrag zur Steigerung der Qualität und Effektivität der Studieneingangsphase und damit zum langfristigen Studienerfolg der Studierenden.

Literatur

Abel, H., & Weber, B. (2014). *28 Jahre Esslinger Modell – Studienanfänger und Mathematik*. In I. Bausch, R. Biehler, R. Bruder, P. R. Fischer, R. Hochmuth, W. Koepf, S. Schreiber & T. Wassong (Hrsg.), *Mathematische Vor- und Brückenkurse* (S. 9–19). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-03065-0_2

Alpers, B. (2013). *A Framework for Mathematics Curricula in Engineering Education: A Report of the Mathematics Working Group*. SEFI.

Assmann, B., & Selke, P. (2010). *Technische Mechanik 1: Band 1 - Statik*. Oldenbourg Wissenschaftsverlag. https://doi.org/10.1524/9783486707939

Ballard, C. L., & Johnson, M. F. (2004). *Basic Math Skills and Performance in an Introductory Economics Class*. *The Journal of Economic Education*, *35*(1), 3–23. https://doi.org/10.3200/JECE.35.1.3-23

Baumert, J., Artelt, C., Klieme, E., & Stanat, P. (2001). *PISA. Programme for International Student Assessment. Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren*. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 285–310). Beltz.

Baumert, J., Bos, W., & Watermann, R. (1999). *TIMSS/III. Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich. Zusammenfassung deskriptiver Ergebnisse* (2. Aufl.). Max-Planck-Institut für Bildungsforschung.

Baumert, J., & Kunter, M. (2006). *Stichwort: Professionelle Kompetenz von Lehrkräften. Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2

Bausch, I., Biehler, R., Bruder, R., Fischer, P. R., Hochmuth, R., Koepf, W., Schreiber, S., & Wassong, T. (Hrsg.). (2014). *Mathematische Vor- und Brückenkurse*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-03065-0

Bean, J. P., & Metzner, B. S. (1985). *A Conceptual Model of Nontraditional Undergraduate Student Attrition. Review of Educational Research*, *55*(4), 485–540. https://doi.org/10.2307/1170245

Beaton, A. E., & Allen, N. L. (1992). *Chapter 6: Interpreting Scales Through Scale Anchoring. Journal of Educational Statistics*, *17*(2), 191–204. https://doi.org/10.3102/10769986017002191

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). *effectsize: Estimation of Effect Size Indices and Standardized Parameters. Journal of Open Source Software*, *5*(56). https://doi.org/10.21105/joss.02815

Bergmann, M., & Franzese, F. (2020). *Fehlende Werte*. In M. Tausendpfund (Hrsg.), *Fortgeschrittene Analyseverfahren in den Sozialwissenschaften* (S. 165–203). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30237-5_6

Blanz, M. (2015). *Forschungsmethoden und Statistik für die Soziale Arbeit: Grundlagen und Anwendungen* (1. Aufl.). Verlag W. Kohlhammer.

Blömeke, S., Schwippert, K., Bremerich-Vos, A., Haudeck, H., Kaiser, G., Nold, G., & Willenberg, H. (Hrsg.). (2011). *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: erste Ergebnisse aus TEDS-LT*. Waxmann.

Blömeke, S., & Zlatkin-Troitschanskaia, O. (2013). *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs (KoKoHs Working Papers, 1)*. https://www.kompetenzen-im-hochschulsektor.de/files/2018/05/KoKoHs_WP1_Bloemeke_Zlatkin-Troitschanskaia_2013_.pdf

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4. Aufl.). Routledge.

Bornkessel, P. (Hrsg.). (2018). *Erfolg im Studium: Konzeptionen, Befunde und Desiderate*. wbv Publikation. https://doi.org/10.3278/6004654w

Borromeo Ferri, R. (2010). *Wege zur Innenwelt des mathematischen Modellierens: Kognitive Analysen zu Modellierungsprozessen im Mathematikunterricht* (1. Aufl.). Vieweg + Teubner. https://doi.org/10.1007/978-3-8348-9784-8

Borromeo Ferri, R., Grünewald, S., & Kaiser, G. (2013). *Effekte kurzzeitiger Interventionen auf die Entwicklung von Modellierungskompetenzen*. In R. Borromeo Ferri, G. Greefrath & G. Kaiser (Hrsg.), *Mathematisches Modellieren für Schule und Hochschule* (S. 41–56). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-01580-0_2

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Springer-Verlag. https://doi.org/10.1007/978-3-642-12770-0

Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T., Jordan, A., Klusmann, U., Tsai, Y.-M., & Neubrand, M. (2006). *Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? Zeitschrift für Erziehungswissenschaft*, *9*(4), 521–544. https://doi.org/10.1007/s11618-006-0166-1

Bundesagentur für Arbeit. (o. J.). *Studiencheck: Wie fit bin ich für mein Studium? - Bundesagentur für Arbeit*. Verfügbar 14. November 2023 unter https://studiencheck.de/

Bundesamt für Migration und Flüchtlinge. (2023). *Das Bundesamt in Zahlen 2022: Asyl, Migration und Integration*. Nürnberg. https://www.bamf.de/SharedDocs/Anlagen/DE/Statistik/BundesamtinZahlen/bundesamt-in-zahlen-2022.pdf?___blob=publicationFile&v=4

Burkhardt, M., Titz, J., & Sedlmeier, P. (2022). *Datenanalyse mit R: fortgeschrittene Verfahren*. Pearson.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed). L. Erlbaum Associates.

Cohen, J. (1992). *A Power Primer. Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cooperation Schule-Hochschule. (2021). *Mindestanforderungskatalog Mathematik Version 3.0 - von Schulen und Hochschulen Baden-Württembergs für ein Studium von WiMINT-Fächern (Wirtschaft, Mathematik, Informatik, Naturwissenschaft und Technik)*. https://cosh-mathe.de/wp-content/uploads/2021/12/makV3.0.pdf

Crouch, C., Fagen, A. P., Callan, J. P., & Mazur, E. (2004). *Classroom demonstrations: Learning tools or entertainment? American Journal of Physics*, *72*(6), 835–838. https://doi.org/10.1119/1.1707018

Dammann, E. (2016). *Entwicklung eines Testinstruments zur Messung fachlicher Kompetenzen in der Technischen Mechanik bei Studierenden ingenieurwissenschaftlicher Studiengänge*. Universität Stuttgart.

Dammann, E., Behrendt, S., Ştefănică, F., & Nickolaus, R. (2016). *Kompetenzniveaus in der ingenieurwissenschaftlichen akademischen Grundbildung – Analysen im Fach Technische Mechanik. Zeitschrift für Erziehungswissenschaft*, *19*(2), 351–374. https://doi.org/10.1007/s11618-016-0675-5

Dammann, E., & Lang, M. (2018). *Mechanisch-mathematisches Modellieren als Prädiktor für Studienerfolg in der Eingangsphase des Bauingenieurstudiums*. In G. Kammasch & J. Petzold (Hrsg.), *Digitalisierung in der Techniklehre: Ihr Beitrag zum Profil Technischer Bildung - 12. Ingenieurpädagogischen Regionaltagung TU Ilmenau vom 11.-13. Mai 2017* (S. 143–148).

Dammann, E., & Lang, M. (2019). *Unterschiede des mathematischen Wissens bei Studierenden des Bauingenieurwesens zu Studienbeginn. Journal of Technical Education (JOTED)*, *7*, 15.

Derr, K., Hübl, R., Mechelke-Schwede, E., Podgayetskaya, T., & Weigel, M. (2017). *Vorhersage von Studienerfolg in den Ingenieurwissenschaften über Learning Analytics? Aussagekraft von Lernerdaten in einem webbasierten Mathematik-Vorkurs*. In U. Kortenkamp & A. Kuzle (Hrsg.), *Beiträge zum Mathematikunterricht 2017 - Vorträge auf der 51. Tagung für Didaktik der Mathematik vom 27.02.2017 bis 03.03.2017 in Potsdam*. WTM-Verlag.

Dittler, U., & Kreidl, C. (2021). *Eine kurze Chronologie der Covid-19-Pandemie im Frühjahr 2020*. In U. Dittler & C. Kreidl (Hrsg.), *Wie Corona die Hochschullehre verändert* (S. 1–13). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32609-8_1

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer. https://doi.org/10.1007/978-3-642-41089-5

Dorsch, F., Wirtz, M. A., & Strohmer, J. (Hrsg.). (2017). *Dorsch - Lexikon der Psychologie* (18. Aufl.). Hogrefe.

Duale Hochschule Baden-Württemberg. (o. J.). *Karriereportal - DHBW Duale Hochschule Baden-Württemberg Mannheim*. Verfügbar 15. November 2023 unter https://www.mathx3.de/quiz/home

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden: mit Online-Materialien* (5. Aufl.). Beltz.

Electric Paper Informationssysteme GmbH. (o. J.). *TeleForm*. Verfügbar 5. Juli 2023 unter https://www.electricpaper.de/produkte/teleform.html

Ernst-Abbe-Hochschule Jena. (o. J.). *Self-Assessment - Maschinenbau B.Eng*. Verfügbar 15. November 2023 unter https://selfassessment.eah-jena.de/osa.php?id=13

Falk, S., & Marschall, M. (2021). *Abbruch des Erststudiums bei MINT-Studierenden: Welche Rolle spielen Informations- und Unterstützungsangebote bei Studienbeginn?* In M. Neugebauer, H.-D. Daniel & A. Wolter (Hrsg.), *Studienerfolg und Studienabbruch* (S. 343–366). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32892-4_15

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). *G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fellenberg, F., & Hannover, B. (2006). *Kaum begonnen, schon zerronnen? Psychologische Ursachenfaktoren für die Neigung von Studienanfängern, das Studium abzubrechen oder das Fach zu wechseln. Empirische Pädagogik*, *20*(4), 381–399.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. Sage.

Fischer, P. R. (2014). *Mathematische Vorkurse im Blended-Learning-Format: Konstruktion, Implementation und wissenschaftliche Evaluation*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-05813-5

Fischer, V. (2019). *Der Einfluss von Interesse und Motivation auf die Messung von Fach- und Bewertungskompetenz im Fach Chemie*. Logos Verlag Berlin.

Fleischer, J., Averbeck, D., Sumfleth, E., Leutner, D., & Brand, M. (2017). *Entwicklung und Vorhersage von Studienzufriedenheit in MINT-Fächern*. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis: Gesellschaft für Didaktik der Chemie und Physik; Jahrestagung in Zürich 2016* (S. 59–62, Bd. *37*). Universität Regensburg. https://doi.org/10.25656/01:12912

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). *Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. Zeitschrift für Erziehungswissenschaft*, *16*(S1), 5–22. https://doi.org/10.1007/s11618-013-0379-z

Fleischer, J., Leutner, D., Brand, M., Fischer, H., Lang, M., Schmiemann, P., & Sumfleth, E. (2019). *Vorhersage des Studienabbruchs in naturwissenschaftlich-technischen Studiengängen. Zeitschrift für Erziehungswissenschaft*, *22*(5), 1077–1097. https://doi.org/10.1007/s11618-019-00909-w

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3. Aufl.). Sage.

Gehlen, C. (2016). *Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie*. Logos Verlag Berlin.

Goedicke, M. (2017). *10 Jahre automatische Bewertung von Programmieraufgaben mit JACK - Rückblick und Ausblick*. https://doi.org/10.18420/IN2017_21

Gold, A. (1988). *Studienabbruch, Abbruchneigung und Studienerfolg: vergleichende Bedingungsanalysen des Studienverlaufs*. P. Lang.

Gollwitzer, M. (2020). *Latent-Class-Analyse (LCA)*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 547–574). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_22

Gross, D., Hauger, W., Schröder, J., & Wall, W. A. (2019). *Technische Mechanik 1: Statik*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-59157-4

Gross, D., Hauger, W., Schröder, J., & Wall, W. A. (2021a). *Technische Mechanik 2: Elastostatik*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61862-2

Gross, D., Hauger, W., Schröder, J., & Wall, W. A. (2021b). *Technische Mechanik 3: Kinetik*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-63065-5

Gross, D., Hauger, W., & Wriggers, P. (2023). *Technische Mechanik 4: Hydromechanik, Elemente der Höheren Mechanik, Numerische Methoden*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-66524-4

HafenCity Universität Hamburg. (o. J.). *Online-Studienorientierung - Bauingenieurwesen*. Verfügbar 15. November 2023 unter http://hcu-studienorientierung.cyquest.net/navigator/bauingenieurwesen/

Hammann, M. (2004). *Kompetenzentwicklungsmodelle. Merkmale und ihre Bedeutung - dargestellt anhand von Kompetenzen beim Experimentieren. Der mathematische und naturwissenschaftliche Unterricht*, *57*(4), 196–203.

Hartig, J. (2007). *Skalierung und Definition von Kompetenzniveaus*. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen: Konzepte*

und Messung - DESI-Studie (Deutsch Englisch Schülerleistungen International) (S. 83–99). Beltz Verlag. https://doi.org/10.25656/01:3143

Hartig, J. (2008). *Kompetenzen als Ergebnisse von Bildungsprozessen*. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Bildungsforschung 26: Kompetenzerfassung in pädagogischen Handlungsfeldern Theorien, Konzepte und Methoden* (S. 15–25). Bundesministerium für Bildung und Forschung (BMBF).

Hartig, J., & Höhler, J. (2010). *Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen. Projekt MIRT*. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung - Zwischenbilanz des DFG- Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (S. 189–198). Beltz Verlag.

Hartig, J., & Klieme, E. (2006). *Kompetenz und Kompetenzdiagnostik*. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Springer-Verlag. https://doi.org/10.1007/3-540-33020-8_9

Hell, B., Linsner, M., & Kurz, G. (2008). *Prognose des Studienerfolgs*. In M. Rentschler (Hrsg.), *Studieneignung und Studierendenauswahl: Untersuchungen und Erfahrungsberichte* (S. 132–177). Shaker Verlag.

Henn, G., & Polaczek, C. (2007). *Studienerfolg in den Ingenieurwissenschaften. Das Hochschulwesen - Forum für Hochschulforschung, -praxis und -politik*, *05/2007*, 144–147.

Heublein, U., Ebert, J., Isleib, S., Hutzsch, C., König, R., Richter, J., & Woisch, A. (2017). *Zwischen Studienerwartungen und Studienwirklichkeit: Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen (Forum Hochschule 01/2017)*. Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH.

Heublein, U., Hutzsch, C., & Schmelzer, R. (2022). *Die Entwicklung der Studienabbruchquoten in Deutschland. DZHW Brief 05/2022*. https://doi.org/10.34878/2022.05.DZHW_BRIEF

Heublein, U., Hutzsch, C., Schreiber, S., Sommer, D., & Besuch, G. (2010). *Ursachen des Studienabbruchs in Bachelor- und in herkömmlichen Studiengängen - Ergebnisse einer bundesweiten Befragung von Exmatrikulierten des Studienjahres 2007/08* (HIS: Forum Hochschule Nr. 02/2010). HIS-Hochschul-Informations-System GmbH. https://www.dzhw.eu/pdf/pub_fh/fh-201002.pdf

Heublein, U., & In der Smitten, S. (2013). *Referenzmodell zur Qualitätssicherung an Fachbereichen und Fakultäten des Maschinenbaus und der Elektrotechnik - Konzept für die Lehre. Maschinenhaus - die VDMA Initiative für Studienerfolg* (HIS-Bericht Nr. 2/4). HIS-Institut für Hochschulforschung. https://vdma.org/documents/34570/17088285/HIS-Bericht+2+-+Konzept+f%C3%BCr+die+Lehre.pdf

Heublein, U., Richter, J., & Schmelzer, R. (2020). *Die Entwicklung der Studienabbruchquoten in Deutschland. DZHW Brief 03/2020.* https://doi.org/10.34878/2020.03.DZHW_BRIEF

Heublein, U., Richter, J., Schmelzer, R., & Sommer, D. (Hrsg.). (2014). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012 (Forum Hochschule 04/2014).* Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH.

Heublein, U., & Schmelzer, R. (2018). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen - Berechnungen auf Basis des Absolventenjahrgangs 2016* (DZHW-Projektbericht Oktober 2018). Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH. https://www.dzhw.eu/pdf/21/studienabbruchquoten_absolventen_2016.pdf

Heublein, U., & Wolter, A. (2011). *Studienabbruch in Deutschland. Definition, Häufigkeit, Ursachen, Maßnahmen. Zeitschrift für Pädagogik*, *57*(2), 214–236. https://doi.org/10.25656/01:8716

Hochschule für Angewandte Wissenschaften Hamburg. (o. J.). *HAW-Navigator: Maschinenbau.* Verfügbar 15. November 2023 unter https://www.haw-navigator.de/mp/

Hochschule für Technik und Wirtschaft Berlin. (o. J.). *Bauingenieurwesen Self-Assessment*. Verfügbar 14. November 2023 unter https://osa.htw-berlin.de/osa.php?id=12

Hochschule für Technik und Wirtschaft des Saarlandes. (o. J.). *Studienorientierung Online*. Verfügbar 14. November 2023 unter https://studienorientierungonline.htwsaar.de/opal/auth/RepositoryEntry/192727939?8

Hochschule Niederrhein. (o. J.). *HSNR-Navigator - Maschinenbau*. Verfügbar 15. November 2023 unter https://navigator.hsnr.de/navigator/maschinenbau/

Hochschule RheinMain. (o. J.). *Physik Vorkurs*. Verfügbar 16. November 2023 unter https://ilias.hs-rm.de/goto.php?target=crs_94599&client_id=DeveloperWiesbaden

Höft, S., & Fischer, T. L. (2023). *Online-Self-Assessments zur Studienorientierung: Überblick zu Modulinhalten und zur Feedbackgestaltung*. In B. Knickrehm, T. Fletemeyer & B.-J. Ertelt (Hrsg.), *Berufliche Orientierung und Beratung* (S. 327–337). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-40601-1_19

Höft, S., Ortner, T., & Hell, B. (2020). *Das D-A-CH-Projekt "OSA-Portal". Übersichtsportal für deutschsprachige Online-Self-Assessments zur studienorientierung und -beratung*. In J. Kohler, P. Pohlenz & U. Schmidt (Hrsg.), *Handbuch Qualität in Studium, Lehre und Forschung. Hauptkapitel E: Methoden und Verfahren des Qualitätsmanagements. Unterkapitel E8: Praxisbeispiele und Innovationen* (E 8.29, S. 93–108). DUZ Verlags- und Medienhaus.

Holland, P. W., & Thayer, D. T. (1988). *Differential Item Performance and the Mantel-Haenszel Procedure DIF IRT*. In H. Wainer & H. I. Braun (Hrsg.), *Test Validity* (S. 129–145). Lawrence Erlbaum Associates Publishers.

Hostenbach, J., & Walpuski, M. (2013). *Untersuchung der Einflussfaktoren auf die Bewertungskompetenz im Fach Chemie. Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 129–157.

IBM. (2022). *IBM SPSS Statistics* (Version 29). https://www.ibm.com/de-de/products/spss-statistics

integral-learning GmbH. (o. J. a). *HM4MINT - NRW*. Verfügbar 16. November 2023 unter https://hm4mint.nrw/hm1/public/index.html

integral-learning GmbH. (o. J. b). *OMB+*. Verfügbar 15. November 2023 unter https://www.ombplus.de/ombplus/public/index.html

Johannes Gutenberg-Universität Mainz. (o. J. a). *KoKoHs (2011-2015): Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor*. Verfügbar 8. November 2023 unter https://www.kompetenzen-im-hochschulsektor.de/kokohs-2011-2015/

Johannes Gutenberg-Universität Mainz. (o. J. b). *KOM-ING: Modellierung und Messung von Kompetenzen der Technischen Mechanik in der Ausbildung von Maschinenbauingenieuren*. Verfügbar 8. November 2023 unter https://www.kompetenzen-im-hochschulsektor.de/kom-ing/

Johannes Gutenberg-Universität Mainz. (o. J. c). *KoM@ING: Kompetenzmodellierungen und Kompetenzentwicklung, integrierte IRT-basierte und qualitative Studien bezogen auf Mathematik und ihre Verwendung im ingenieurwissenschaftlichen Studium*. Verfügbar 8. November 2023 unter https://www.kompetenzen-im-hochschulsektor.de/koming/

Johannes Gutenberg-Universität Mainz. (o. J. d). *MoKoMasch: Modellierung von Kompetenzen bei Studierenden des Maschinenbaus in den Bereichen Konstruktion, Entwurf und Produktionstechnik*. https://www.kompetenzen-im-hochschulsektor.de/mokomasch/

Kassambara, A. (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. https://CRAN.R-project.org/package=rstatix

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. Logos Verlag Berlin.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). *Standardbezogene kompetenzmodellierung in den naturwissenschaften der sekundarstufe. Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 135–153.

Kelava, A., & Moosbrugger, H. (2020). *Einführung in die Item-Response-Theorie (IRT)*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheo-*

rie und Fragebogenkonstruktion (S. 369–409). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_16

Kelava, A., Noventa, S., & Robitzsch, A. (2020). *Überblick über Modelle der Item-Response-Theorie (IRT)*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 425–446). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_18

Kempen, L., & Wassong, T. (2017). *VEMINT mobile with Apps: der gezielte Einsatz von mobilen Endgeräten in einem Mathematik- Vorkurs unter Verwendung der multimedialen VEMINT-Materialien*. In R. Kordts-Freudinger, D. Al-Kabbani & N. Schaper (Hrsg.), *Hochschuldidaktik im Dialog - Beiträge der Jahrestagung der Deutschen Gesellschaft für Hochschuldidaktik (dghd) 2015* (S. 13–38). wbv.

Key, O., & Hill, L. (2018). *Modellansätze ausgewählter Hochschulen zur Neugestaltung der Studieneingangsphase*. Hochschulrektorenkonferenz. https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-02-Publikationen/CHE_07032018_final.pdf

Klieme, E. (2004). *Was sind Kompetenzen und wie lassen sie sich messen? Pädagogik*, *56*(6), 10–13.

Klieme, E., & Leutner, D. (2006). *Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG*. https://doi.org/10.25656/01:4493

Klöpping, S., Scherfer, M., Gokus, S., Dachsberger, S., Krieg, A., Wolter, A., Bruder, R., Ressel, W., & Umbach, E. (Hrsg.). (2017). *Studienabbruch in den Ingenieurwissenschaften. Empirische Analyse und Best Practices zum Studienerfolg (acatech STUDIE)*. Herbert Utz Verlag.

Kobow, I. (2015). *Entwicklung und Validierung eines Testinstrumentes zur Erfassung der Kommunikationskompetenz im Fach Chemie*. Logos Verlag Berlin.

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012, 12. September). *Das Rasch Modell in der Praxis: Eine Einführung in eRm* (1. Aufl.). utb GmbH. https://doi.org/10.36198/9783838537863

Koller, I., Maier, M. J., & Hatzinger, R. (2015). *An Empirical Power Analysis of Quasi-Exact Tests for the Rasch Model: Measurement Invariance in Small Samples*. *Methodology*, *11*(2), 45–54. https://doi.org/10.1027/1614-2241/a000090

Köller, O., Baumert, J., & Bos, W. (2001). *TIMSS - Third International Mathematics and Science Study. Dritte internationale Mathematik- und Naturwissenschaftsstudie*. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 269–284). Beltz.

Köller, O., Pant, H. A., & Zlatkin-Troitschanskaia, O. (2020). *Diagnostik von studentischen Kompetenzen im Hochschulsektor: Einleitung in den Thementeil*. *Diagnostica*, *66*(2), 77–79. https://doi.org/10.1026/0012-1924/a000252

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). *MVN: An R Package for Assessing Multivariate Normality*. *The R Journal*, *6*(2), 151–162. https://doi.org/10.32614/RJ-2014-031

Kroher, M., Beuße, M., Becker, K., Ehrhardt, M.-C., Koopmann, J., Schommer, T., Schwabe, U., Steinkühler, J., Völk, D., Peter, F., & Buchholz, S. (2023). *Die Studierendenbefragung in Deutschland: 22. Sozialerhebung - Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2021*. Bundesministerium für Bildung und Forschung (BMBF). https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/4/31790_22_Sozialerhebung_2021.pdf

Kultusministerkonferenz. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Carl Link.

Lehmann, M. (2018). *Relevante mathematische Kompetenzen von Ingenieurstudierenden im ersten Studienjahr - Ergebnisse einer empirischen Untersuchung*. Humboldt-Universität zu Berlin.

Leutner, D., Fleischer, J., Grünkorn, J., & Klieme, E. (Hrsg.). (2017). *Competence Assessment in Education*. Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0

LimeSurvey GmbH. (o. J.). *LimeSurvey — Free Online Survey Tool*. Verfügbar 12. Dezember 2023 unter https://www.limesurvey.org/de

Little, R. J. A. (1988). *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Lord, F. M. (1980). *Applications of Item Response Theory To Practical Testing Problems*. Routledge. https://doi.org/10.4324/9780203056615

Magnus, K., & Müller-Slany, H. H. (2009). *Grundlagen der Technischen Mechanik* (7. Aufl.). Springer Fachmedien GmbH. https://doi.org/10.1007/978-3-663-01626-7

Martschink, B. (2013). *Open Educational Resources in Einführungsveranstaltungen der Ingenieurmathematik*. *Zeitschrift für Hochschulentwicklung*, *8*(4), 1–11. https://doi.org/10.3217/zfhe-8-04/02

Middendorff, E., Apolinarski, B., Becker, K., Bornkessel, P., Brandt, T., Heißenberg, S., & Poskowsky, J. (2017). *Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2016: Zusammenfassung zur 21. Sozialerhebung des Deutschen Studentenwerks - durchgeführt vom Deutschen Zentrum für Hochschul- und Wissenschaftsforschung*. Bundesministerium für Bildung und Forschung (BMBF). Berlin. https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/4/31338_21_Sozialerhebung_2016_Zusammenfassung.pdf

Ministeriums für Schule und Bildung NRW. (2021). *Zentrale Vergleichsarbeiten (Lernstandserhebungen)*. Verfügbar 2. November 2023 unter https://bass.schul-welt.de/6912.htm

Ministeriums für Schule und Bildung NRW. (2022). *Kernlehrplan für die Sekundarstufe II Gymnasium/Gesamtschule in Nordrhein-Westfalen - Physik*. https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan/332/gost_klp_ph_2022_06_07.pdf

Ministeriums für Schule und Bildung NRW. (2023). *Kernlehrplan für die Sekundarstufe II Gymnasium/Gesamtschule in Nordrhein-*

Westfalen - Mathematik. https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan/331/gost_klp_m_2023_06_07.pdf

MINT-Kolleg Baden-Württemberg. (o. J. a). *Online-Brückenkurs Mathematik*. Verfügbar 16. November 2023 unter https://lx3.mint-kolleg.kit.edu/onlinekursmathev4/html/sectionx3.1.0.html

MINT-Kolleg Baden-Württemberg. (o. J. b). *Online-Brückenkurs Physik*. Verfügbar 16. November 2023 unter https://lx3.mint-kolleg.kit.edu/onlinekursphysik/html/sectionx3.1.0.html

Moodle Pty Ltd. (o. J.). *Moodle.org*. Verfügbar 19. November 2023 unter https://moodle.org/

Moosbrugger, H., & Brandt, H. (2020). *Antwortformate und Itemtypen*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 91–117). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_5

Moosbrugger, H., Schermelleh-Engel, K., Gäde, J. C., & Kelava, A. (2020). *Testtheorien im Überblick*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 251–273). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_12

Müller, J., Stender, A., Fleischer, J., Borowski, A., Dammann, E., Lang, M., & Fischer, H. E. (2018). *Mathematisches Wissen von Studienanfängern und Studienerfolg. Zeitschrift für Didaktik der Naturwissenschaften*, *24*(1), 183–199. https://doi.org/10.1007/s40573-018-0082-y

Müller, K., Prenzel, M., Sälzer, C., Mang, J., Heine, J.-H., & Gebhardt, M. (2017). *Wie schneiden Schülerinnen und Schüler an Sonder- und Förderschulen bei PISA ab? - Analysen aus der PISA 2012-Zusatzerhebung zu Jugendlichen mit sonderpädagogischem Förderbedarf. Unterrichtswissenschaft*, *45*(2), 175–192. https://epub.uni-regensburg.de/43713/

Müller-Slany, H. H. (2018). *Aufgaben und Lösungsmethodik Technische Mechanik: mit Strategie Lösungen systematisch erarbeiten* (2. Aufl.). Springer Vieweg.

Narciss, S. (2006). *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse*. Waxmann.

Neumann, K. (2020). *Die Bedeutung instruktionaler Kohärenz für eine systematische Kompetenzentwicklung. Unterrichtswissenschaft*, *48*(1), 1–10. https://doi.org/10.1007/s42010-020-00068-6

Nolden, P. (2019). *Studentisches Erleben und Studienabbruchneigung : Entwicklung und Überprüfung eines multikausalen und multiperspektivischen Erklärungsmodells im Hochschulkontext*. Rheinisch-Westfälischen Technischen Hochschule Aachen.

Oberman, H. (2023). *ggmice: Visualizations for 'mice' with 'ggplot2'*. https://CRAN.R-project.org/package=ggmice

OECD. (2004). *Lernen für die Welt von morgen: erste Ergebnisse von PISA 2003*. Elsevier, Spektrum, Akad. Verl.

Organisation for Economic Co-operation and Development Organisation. (2023). *PISA 2025 Science Framework (Second Draft)*. Paris. https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf

OSA-Portal. (o. J.). *OSA-Portal - Das unabhängige Vergleichsportal für Online Self Assessments zur Studienorientierung*. Verfügbar 15. November 2023 unter https://www.osa-portal.de/index.php

Pelz, M., Lang, M., Özmen, Y., Schröder, J., Walker, F., & Müller, R. (2020). *Kann der individuelle Lernerfolg von Studierenden der Bauwissenschaften durch interaktive online Module gefördert werden? [Tagungsvortrag]*. dghd20 (49. Jahrestagung der Deutschen Gesellschaft für Hochschuldidaktik), Freie Universität Berlin, Tagung wurde abgesagt. https://doi.org/10.13140/RG.2.2.35969.40802

Pelz, M., Lang, M., Özmen, Y., Schröder, J., Walker, F., & Müller, R. (2021). *Online-Selbsttests und -Vorkurse - ein Mittel zur nachhaltigen Qualitätsentwicklung im Bauingenieurwesen?* In A. Thielsch, C. Bade & L. Mitterauer (Hrsg.), *Ursprünge hinterfragen - Vielfalt ergründen - Praxis einordnen (ReGeneration Hochschullehre)* (S. 97–109). wbv Publikation.

Pelz, M., Lang, M., Walker, F., Özmen, Y., Schröder, J., & Müller, R. (2019). *Verstetigung digitaler Hochschullehre in der Technischen Mechanik Adaption eines Referenzmodells zur Qualitätssicherung in den Studienverlauf der Bauwissenschaften*. In S. Heuchemer, B. Szczyrba & S. Spöth (Hrsg.), *Hochschuldidaktik erforscht Qualität: Forschung und Innovation in der Hochschulbildung III* (S. 97–111). Cologne Open Science.

Prusty, B. G., Ho, O., & Ho, S. (2009). *Adaptive Tutorials using eLearning Platform for Solid Mechanics Course in Engineering. Proceedings of the AAEE 2009 20th Annual Conference*, 828–833.

Prusty, B. G., & Russell, C. (2011). *Engaging students in learning threshold concepts in engineering mechanics: adaptive eLearning tutorials. Proceedings of the 17th International Conference on Engineering Education (ICEE)*.

Prusty, G., Russell, C., Ford, R., Ben-Naim, D., Ho, S., Vrcelj, Z., Marcus, N., McCarthy, T., Goldfinch, T., Ojeda, R., Gardner, A., Molyneaux, T., & Hadgraft, R. (2011). *Adaptive tutorials to target Threshold Concepts in Mechanics – a community of practice approach. Proceedings of the 2011 AaeE Conference*, 305–311.

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.1). Vienna, Austria. https://www.R-project.org/

Raju, N. S. (1990). *Determining the Significance of Estimated Signed and Unsigned Areas Between Two Item Response Functions. Applied Psychological Measurement*, *14*(2), 197–207. https://doi.org/10.1177/014662169001400208

Rauch, D., & Hartig, J. (2020). *Interpretation von Testwerten in der Item-Response-Theorie (IRT)*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 411–424). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_17

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. https://CRAN.R-project.org/package=psych

Rheinisch-Westfälischen Technischen Hochschule Aachen. (o. J.). *SelfAssessment Bauingenieurwesen*. Verfügbar 15. November 2023 unter https://assess.rwth-aachen.de/tm4_rwth/frontend/www/index.php?r=uberTest/view&id=3

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM

Ropohl, M. J. (2010). *Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion: Entwicklung und Analyse von Testaufgaben*. Logos Verlag Berlin.

Rose, N. (2020). *Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT)*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 447–500). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_19

Rosseel, Y. (2012). *lavaan: An R Package for Structural Equation Modeling*. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2. Aufl.). H. Huber.

Schiepe-Tiska, A., Rönnebeck, S., & Neumann, K. (2019). *Naturwissenschaftliche Kompetenz in PISA 2018 – aktueller Stand, Veränderungen und Implikationen für die naturwissenschaftliche Bildung in Deutschland*. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018. Grundbildung im internationalen Vergleich* (1. Aufl., S. 211–240). Waxmann.

Schmidt-Atzert, L., Amelang, M., Fydrich, T., Moosbrugger, H., & Zielinski, W. (2012). *Psychologische Diagnostik*. Springer-Verlag Berlin Heidelberg.

Schmitt, Niermann, Knutzen & Klaffke. (2018). *Betrachtung der Studiensituation durch den Einsatz einer Online-Selbsteinschätzung*. *Zeitschrift für Hochschulentwicklung*, *13*(3), 113–130. https://doi.org/10.3217/zfhe-13-03/07

Schoppmeier, F. (2013). *Physikkompetenz in der gymnasialen Oberstufe: Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*. Logos Verlag Berlin.

Schröder, J. (2020). *Vorlesungsskript - Modul Mechanik 1 (Stereostatik)*.

Schröder, J. (2021). *Vorlesungsskript - Modul Mechanik 2 (Elastostatik)*.

Schworm, S. (2004). *Lernen aus Beispielen: computerbasierte Lernumgebungen zum Erwerb argumentativer und didaktischer Fertigkeiten*. Albert-Ludwigs-Universität Freiburg i.Br.

Sedlmeier, P., & Burkhardt, M. (2021). *Datenanalyse mit R: Beschreiben, Explorieren, Schätzen und Testen*. Pearson.

Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (3. Aufl.). Pearson.

Shapiro, S. S., & Wilk, M. B. (1965). *An Analysis of Variance Test for Normality (Complete Samples)*. *Biometrika*, *52*(3/4), 591. https://doi.org/10.2307/2333709

Spady, W. G. (1970). *Dropouts from Higher Education: An Interdisciplinary Review and Synthesis*. *Interchange*, *1*, 64–85. https://doi.org/10.1007/BF02214313

Statistisches Bundesamt. (2023). *Genesis-Online*. Verfügbar 2. November 2023 unter https://www-genesis.destatis.de/genesis/online

Streiner, D. L. (2003). *Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency*. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Strobl, C. (2012). *Das Rasch-Modell: eine verständliche Einführung für Studium und Praxis* (2. Aufl.). Hampp.

Ströhlein, G. (1983). *Bedingungen des Studienabbruchs: Eine längsschnittuntersuchung bei Studenten ingenieurwissenschaftlicher Fakultäten*. Lang.

Swaminathan, H., & Rogers, H. J. (1990). *Detecting Differential Item Functioning Using Logistic Regression Procedures*. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Technische Hochschule Würzburg-Schweinfurt. (o. J.). *OSA Bauingenieurwesen*. Verfügbar 15. November 2023 unter https://orientierung.fhws.de/mod/scorm/view.php?id=137

Technische Universität Dresden. (o. J.). *Online-Vorbereitungskurs Physik*. Verfügbar 16. November 2023 unter https://bildungsportal.sachsen.de/opal/auth/RepositoryEntry/13492649994/

Technische Universität Hamburg. (o. J.). *mytrack | Online-Selbsteinschätzung*. Verfügbar 14. November 2023 unter https://mytrack-tuhh.de/angebote/online-selbsteinschaetzung

Technische Universität München. (o. J. a). *PISA 2025: Lernen in der digitalen Welt*. Verfügbar 2. November 2023 unter https://www.pisa.tum.de/pisa/kompetenzbereiche/lernen-in-der-digitalen-welt/

Technische Universität München. (o. J. b). *PISA 2025: Naturwissenschaftliche Kompetenz*. Verfügbar 2. November 2023 unter https://www.pisa.tum.de/pisa/kompetenzbereiche/naturwissenschaftliche-kompetenz/

Thiele, L., & Kauffeld, S. (2019). *Online Self-Assessments zur Studien- und Universitätswahl*. In S. Kauffeld & D. Spurk (Hrsg.), *Handbuch Karriere und Laufbahnmanagement* (S. 109–132). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48750-1_4

Tieben, N. (2019). *Brückenkursteilnahme und Studienabbruch in Ingenieurwissenschaftlichen Studiengängen. Zeitschrift für Erziehungswissenschaft*, *22*(5), 1175–1202. https://doi.org/10.1007/s11618-019-00906-z

TIMSS & PIRLS International Study Center. (o. J.). *TIMSS: Trends in International Mathematics and Science Study*. Verfügbar 31. Oktober 2023 unter https://timssandpirls.bc.edu/timss-landing.html

Tinto, V. (1975). *Dropout from Higher Education: A Theoretical Synthesis of Recent Research. Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Pub. Co.

Universität Kassel. (o. J.). *Online-Studienwahlassistent (OSA) - Bachelor Bauingenieurwesen*. Verfügbar 15. November 2023 unter https://www.uni-kassel.de/osa-fb14bau/bauingenieurwesen/

Universität Paderborn. (o. J.). *studiVEMINT*. Verfügbar 15. November 2023 unter https://fddm.uni-paderborn.de/projekte/studivemint/allgemeines/

Urban, D., & Mayerl, J. (2018). *Angewandte Regressionsanalyse: Theorie, Technik und Praxis*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-01915-0

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

VDI Verein Deutscher Ingenieure e.V. & Institut der Deutschen Wirtschaft e.V. (2023). *Der regionale Arbeitsmarkt in den Ingenieurberufen. Sonderteil: Weibliche Ingenieurbeschäftigte* (Ingenieurmonitor Nr. 2023/I). https://www.vdi.de/ueber-uns/presse/publikationen/details/vdi-iw-ingenieurmonitor-1-quartal-2023

Vink, G., Frank, L. E., Pannekoek, J., & Van Buuren, S. (2014). *Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica*, *68*(1), 61–90. https://doi.org/10.1111/stan.12023

Voßkamp, R., & Laging, A. (2014). *Teilnahmeentscheidungen und Erfolg: Eine Fallstudie zu einem Vorkurs aus dem Bereich der Wirtschaftswissenschaften*. In I. Bausch, R. Biehler, R. Bruder, P. R. Fischer, R. Hochmuth, W. Koepf, S. Schreiber & T. Wassong (Hrsg.), *Mathematische Vor- und Brückenkurse* (S. 67–83). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-03065-0_6

Walpuski, M., Kauertz, A., Kampa, N., Fischer, H. E., Mayer, J., Sumfleth, E., & Wellnitz, N. (2010). *ESNaS - Evaluation der Standards für die Naturwissenschaften in der Sekundarstufe I*. In A. Gehrmann, U. Hericks & M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle: Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 171–184). Klinkhardt.

Weinert, F. E. (1999). *Concepts of Competence*. OECD. Paris.

Weinert, F. E. (2001). *Concept of Competence: A Conceptual Clarification*. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Hogrefe & Huber.

Weinert, F. E. (2014). *Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit*. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (3. Aufl., S. 17–31). Beltz.

Weis, M., & Reiss, K. (2019). *PISA 2018 - Ziele und Inhalte der Studie*. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018: Grundbildung im internationalen Vergleich* (S. 13–20). Waxmann Verlag GmbH.

Wess, R., Klock, H., Siller, H.-S., & Greefrath, G. (2021). *Test Quality*. In *Measuring Professional Competence for the Teaching of Mathematical Modelling* (S. 77–84). Springer International Publishing. https://doi.org/10.1007/978-3-030-78071-5_4

Willige, J., Woisch, A., Grützmacher, J., & Naumann, H. (2014). *Studienqualitätsmonitor SQM 2014: Online-Befragung Studierender im Sommersemester 2014 (Fächergruppen an Universitäten)*. Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH.

Wilson, M. (2003). *On Choosing a Model for Measuring. Methods of Psychological Research*, *8*(3), 1–22. https://doi.org/10.23668/psycharchives.12789

Wilson, M. (2023). *Constructing Measures: An Item Response Modeling Approach* (2. Aufl.). Routledge. https://doi.org/10.4324/9781003286929

Wirtz, M. A. (2004). *Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann. Die Rehabilitation*, *43*(2), 109–115. https://doi.org/10.1055/s-2003-814839

Zlatkin-Troitschanskaia, O., & Kuhn, C. (2010). *Messung akademisch vermittelter Fertigkeiten und Kenntnisse von Studierenden bzw. Hochschulabsolventen: Analyse zum Forschungsstand*.

Zlatkin-Troitschanskaia, O., Pant, H. A., Nagel, M.-T., Molerov, D., Lautenbach, C., & Toepper, M. (Hrsg.). (2020). *KoKoHs Assessment-*

Portfolio: Testverfahren zur Modellierung und Messung generischer und domänenspezifischer Kompetenzen bei Studierenden und Hochschulabsolventen. pfalzdruck.

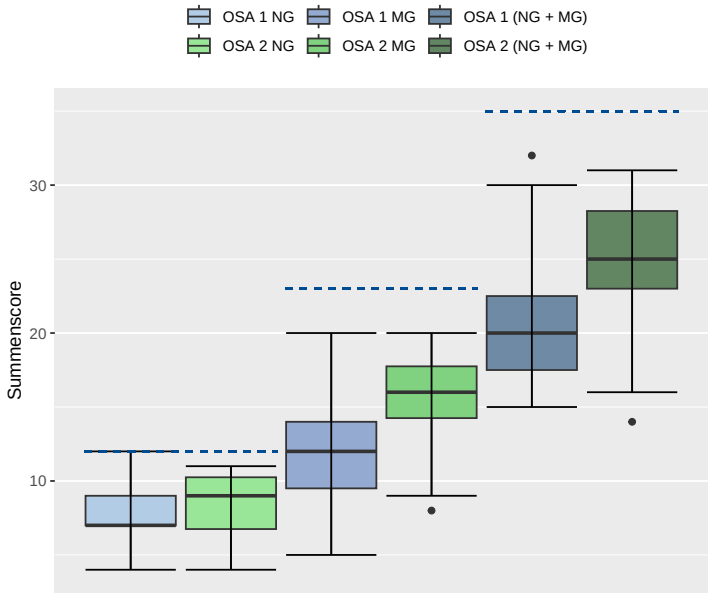
Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., Braunheim, D., & Molerov, D. (2021). *KoKoHs-Map: Landkarte zum Kompetenzerwerb im Hochschulbereich und den Einflussfaktoren. Eine Metastudie zu den Ergebnissen der KoKoHs-Förderlinie (2011-2020)*. https://www.kompetenzen-im-hochschulsektor.de/files/2021/08/Zlatkin-Troitschanskaia-et-al_2021_KoKoHs-Map_Landkarte-zum-Kompetenzerwerb-im-Hochschulbereich_final.pdf

Anhang A

Abbildungen: Extremwertanalyse - Box-Whisker-Diagramme (PMM)

Abbildung A.1

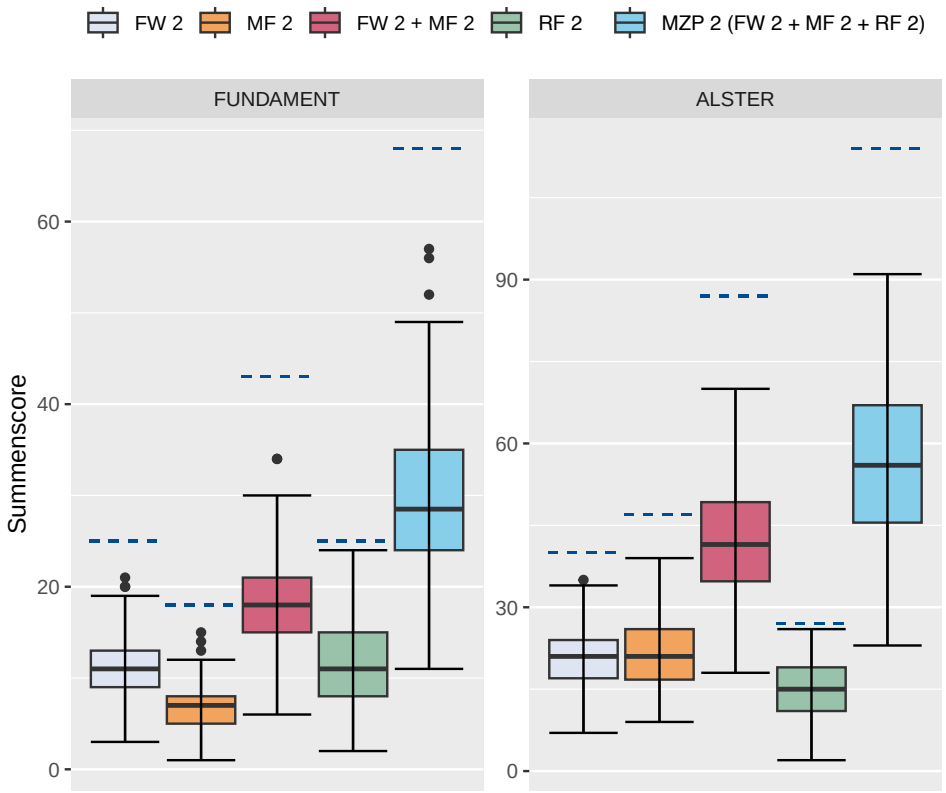
Box-Whisker-Diagramm der Summenscores an MZP 1 (PMM)



Anmerkungen. MZP = Messzeitpunkt; OSA = Online-Self-Assessment; NG = naturwissenschaftliche Grundlagen; MG = mathematische Grundlagen.

Abbildung A.2

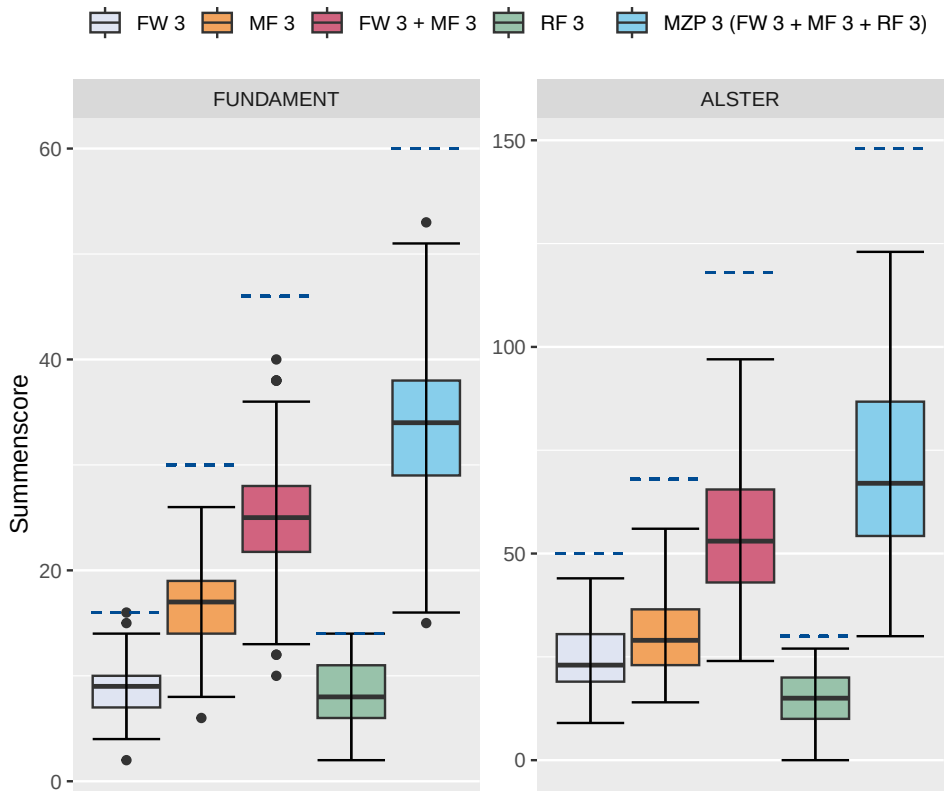
Box-Whisker-Diagramm der Summenscores an MZP 2 (PMM)



Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; MZP = Messzeitpunkt.

Abbildung A.3

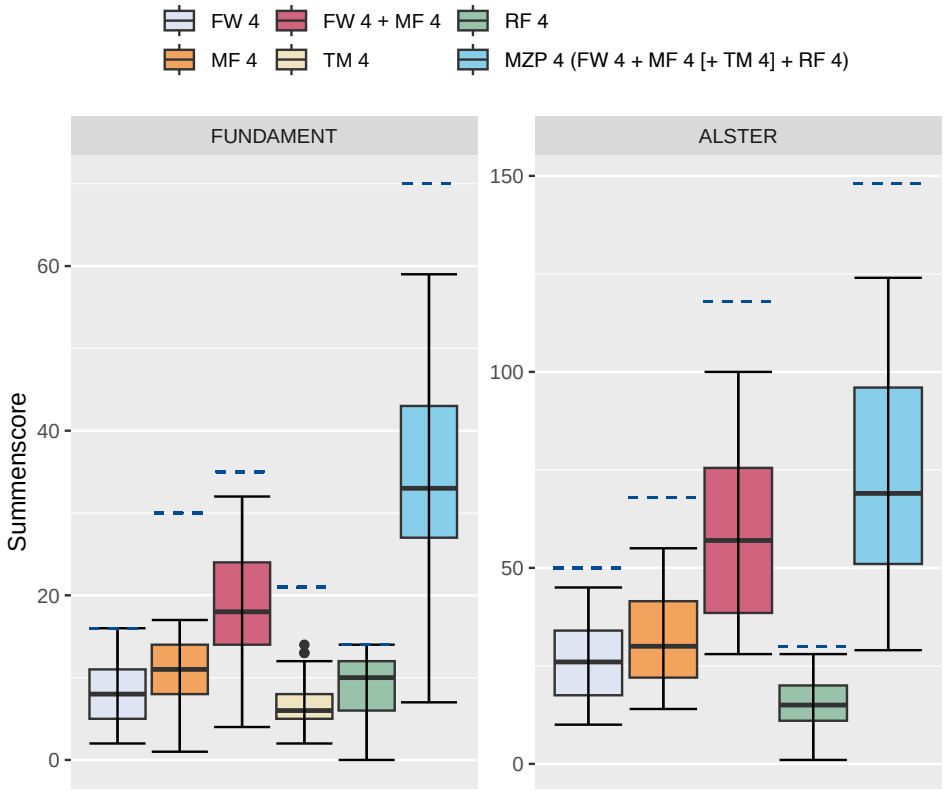
Box-Whisker-Diagramm der Summenscores an MZP 3 (PMM)



Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; MZP = Messzeitpunkt.

Abbildung A.4

Box-Whisker-Diagramm der Summenscores an MZP 4 (PMM)



Anmerkungen. FW = Fachwissen; MF = Modellierungsfähigkeit; RF = Rechenfähigkeit; TM_4 = Technische Mechanik 2-Items an MZP 4; MZP = Messzeitpunkt.