



TUM School of Education
Professur für Fachdidaktik Life Sciences

Biologiespezifisches Bildverständnis beim Umgang mit multiplen externen Repräsentationen

Entwicklung und Validierung eines Kompetenzmodells zur Integration von
Diagrammen und Schemata

Christina Heidi Beck

Vollständiger Abdruck der von der Fakultät TUM School of Education
der Technischen Universität München zur Erlangung des akademischen Grades
eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Peter Hubwieser
Prüfende der Dissertation:	1. Prof. Dr. Claudia Nerdel 2. Prof. Dr. Julia Schwanewedel

Die Dissertation wurde am 13.07.2017 bei der Technischen Universität München
eingereicht und durch die Fakultät TUM School of Education am 27.09.2017 ange-
nommen.

Die Wissenschaft fängt eigentlich erst da an, interessant zu werden, wo sie aufhört.

- Justus von Liebig -

INHALTSVERZEICHNIS

Zusammenfassung	VI
I Einleitung.....	1
II Theoretischer Hintergrund.....	4
1 Naturwissenschaftliche Kommunikation	4
1.1 Scientific Literacy	4
1.2 Kommunikative Kompetenzen in den Bildungsstandards	7
1.3 Fachsprache	9
1.4 Einteilung von Repräsentationen.....	12
1.4.1 Texte.....	13
1.4.2 Bilder	14
1.4.2.1 Realistische Bilder	16
1.4.2.2 Schemazeichnungen.....	17
1.4.2.3 Logische Bilder	19
1.4.3 Text-Bild-Kombinationen, MER	23
1.5 Biologiespezifische Repräsentationen und Kontext	31
1.6 Zusammenfassung.....	33
2 Theorien und Modelle zum Text-Bild-Verstehen	35
2.1 Repräsentationskompetenz.....	35
2.2 Kognitionspsychologischer Hintergrund.....	38
2.3 Multimedia-Learning	40
2.4 Text-Bild-Integrationsmodell	43
2.5 Zusammenfassung.....	47
3 Kompetenzmodell zur Integration von MER.....	48
3.1 Kompetenzbegriff.....	48

3.2 Kompetenzmodellierung.....	49
3.2.1 Kompetenzstrukturmodelle	50
3.2.2 Kompetenzniveauamodelle	52
3.2.3 Kompetenzentwicklungsmodelle	52
3.3 Kompetenzstrukturmodell zur Integration von MER	53
3.3.1 Dimension Umgang mit multiplen externen Repräsentationen.....	55
3.3.2 Dimension Aufgabenkontext	57
3.3.3 Dimension Repräsentationstyp	58
3.4 Zusammenfassung	60
III Ziele, Forschungsfragen und Hypothesen	61
IV Empirischer Teil.....	67
4 Material der Vorstudie	67
4.1 Stichprobe.....	67
4.2 Testkonstruktion.....	68
4.2.1 Operationalisierung	71
4.2.2 Aufgabenkonstruktion	72
4.3 Testheftdesign.....	82
4.4 Durchführung der Vorstudie	86
5 Statistische Methoden der Vorstudie	87
5.1 Klassische und Probabilistische Testtheorie.....	87
5.2 Grundlagen der probabilistischen Testtheorie	89
5.2.1 Das eindimensionale Rasch-Modell	90
5.2.2 Schätzung von Item- und Personenparametern.....	92
5.3 Auswertungsmethoden der Vorstudie	94
5.3.1 Vorbereitende Schritte	94
5.3.2 Qualitative Inhaltsanalyse.....	94

5.3.3 Analyse der Itemparameter	98
5.3.4 Deskriptive Analysen im Rahmen der Klassischen Testtheorie	102
5.4 Gütekriterien.....	105
6 Ergebnisse der Vorstudie	107
6.1 Kategoriensystem für offene Aufgaben	107
6.2 Rasch-Skalierung: interne Konsistenz des Kompetenztests.....	112
6.3 Itemselektion.....	118
6.4 Schwierigkeitsunterschiede.....	119
6.5 Zusammenfassung und Implikationen für die Hauptstudie.....	125
7 Material der Hauptstudie.....	128
7.1 Stichprobe und Organisation der Testung.....	128
7.2 Testkonstruktion.....	129
7.3 Testheftdesign.....	132
8 Statistische Methoden der Hauptstudie.....	134
8.1 Probabilistische Testtheorie für Mehrdimensionale Modelle.....	134
8.1.1 Das mehrdimensionale Rasch-Modell	134
8.1.2 Schätzung von Item- und Personenparametern.....	137
8.1.3 Messmodelle der Untersuchung	138
8.2 Prüfung der Modellpassung	142
8.2.1 Prüfung der Reliabilität.....	142
8.2.2 Prüfung der Modellgültigkeit und Modelldimensionen	143
8.3 Weitere Auswertungsmethoden der Hauptstudie	146
8.3.1 Analyse der Item- und Personenparameter.....	146
8.3.2 Korrelationsanalysen.....	147
8.3.3 Regressionsanalysen	148

9 Ergebnisse der Hauptstudie	152
9.1 Testitems und Modellpassung.....	152
9.1.1 Das 1-dimensionale Modell zur Repräsentationskompetenz.....	152
9.1.2 Das 3-dimensionale Modell zur MER-Integration.....	155
9.1.3 Das 3-dimensionale Kontextmodell zur MER-Integration.....	157
9.1.4 Das 2-dimensionale Repräsentationsmodell zur MER-Integration.....	160
9.1.5 Das 2-dimensionale Modell zur MER-Integration.....	162
9.1.6 Das 6-dimensionale Within-Modell zur MER-Integration.....	164
9.1.7 Zusammenfassung der Modellpassung.....	164
9.2 Modellvergleich.....	165
9.3 Schwierigkeitsunterschiede und Vorhersage der Schwierigkeit.....	170
9.3.1 MER-Integration.....	171
9.3.2 Aufgabenkontext und Repräsentationstyp.....	174
9.3.3 Schwierigkeitsgenerierende Merkmale.....	175
9.4 Prüfung von Kompetenzunterschieden.....	179
9.4.1 Kompetenzausprägungen der Teilkompetenzen im Vergleich.....	179
9.4.2 Studiengang.....	182
9.4.3 Noten, Abiturfach und Geschlecht.....	185
9.5 Qualitative Fehleranalyse.....	186
9.5.1 Informationsentnahmenähe Integration (MERI).....	187
9.5.2 Konstruktionsnähe Integration (MERII).....	188
9.5.3 Translation und Transformation aus mehreren Repräsentationen (MERIII).....	192
V Diskussion	195
10 Methodische Diskussion	195
10.1 Testentwicklung, Stichprobe und Validierung.....	195
10.2 Rasch-Skalierung und Testdesign.....	198

11 Inhaltliche Diskussion.....	203
11.1 Modellvergleich und Modellstruktur	203
11.2 Differenzierung und Ausprägung der Kompetenzen	206
11.3 Kompetenz, Studiengang, Noten und Geschlecht.....	209
12 Relevanz und Ausblick	213
VI Verzeichnisse.....	221
13.1 Literaturverzeichnis	221
13.2 Tabellenverzeichnis	251
13.3 Abbildungsverzeichnis	254
Anhang	257
14. 1 Testinstrumente	257
14.1.1 Itembeispiele der Hauptstudie	257
14.1.2 Testheftdesign Pilotierung	267
14.1.3 Testheftdesign Hauptstudie	268
14.2 Ergänzungen zur Ergebnisdarstellung	271
14.3 Publikationen.....	275
14.3.1 Referierte Beiträge zu Forschungsthemen	275
14.3.2 Beiträge in Abstractbänden	275
14.3.3 Tagungspräsentationen	276
Danksagung	278

ZUSAMMENFASSUNG

Naturwissenschaftliche Kommunikation bedeutet fachspezifische Kommunikation unter Verwendung unterschiedlicher Repräsentationen, die uns in Alltag, Wissenschaft und Unterricht begegnen. Für den gesellschaftlichen Diskurs nimmt die Beherrschung naturwissenschaftlicher Fachsprache eine Schlüsselstellung ein.

„Communication skills are one of the most important competencies for 21st century global citizens” (Chung et al., 2016, S. 1).

Fähigkeiten im Umgang mit unterschiedlichen Darstellungsformen werden entscheidend für eine naturwissenschaftliche Grundbildung im Rahmen von Scientific Literacy.

„Visualization skills, reasoning skills, tree building skills, representational competence and systems thinking skills a part of science education in general in the 21st century” (Treagust & Tsui, 2013, S. 17)

Dabei sind Lernende mit einer enormen Vielfalt an Darstellungen konfrontiert, wobei Text-Bild-Kombinationen in Lehr-Lernmaterialien wie z.B. in Schulbüchern eher die Regel als die Ausnahme sind. Der Umgang mit unterschiedlichen Darstellungsformen wird so zu einem konstituierenden Merkmal fachsprachlicher Kompetenz und entscheidend bei der Vermittlung von Kommunikationsfähigkeiten im naturwissenschaftlichen Unterricht.

Ziel der Dissertation war die Entwicklung eines Kompetenzstrukturmodells, das Kompetenzen von Studienanfängerinnen und Studienanfängern zum Umgang mit multiplen externen Repräsentationen (MER) im Fach Biologie abbildet. Das entwickelte Kompetenzmodell berücksichtigt drei Modellkomponenten: *Informationsentnahmenahe Integration (MERI)*, *Konstruktionsnahe Integration (MERII)*, *Translation und Transformation aus mehreren Repräsentationen (MERIII)*, sowie verschiedene *Aufgabenkontexte* der Biologie (*Ökologie, Stoffwechsel, Genetik*) und zwei *Repräsentationstypen* (*Diagramm, Schemazeichnung*). Über Testaufgaben wurde das theoretisch hergeleitete Konstrukt operationalisiert und erfasst. Anschließend wurde das postulierte Kompetenzmodell auf seine empirische Gültigkeit durch Anwendung qualitativer und quantitativer Testverfahren überprüft. Dieser Prozess ermöglicht die Diagnose einer personenbezogenen Repräsentationskompetenz sowie das Identifizieren schwierigkeitsgenerierender Aufgabenmerkmale. Aus 210 Testaufgaben der Pilotierung wurden in der Hauptstudie 36 Testaufgaben im offenen und geschlossenen („2 aus 5“) Antwortformat in einem Youden-Square-Design, bestehend aus unterschiedlichen Testheften, den Probanden zur Beantwortung vorgelegt. An der Untersuchung nahmen Studierende (N=548)

der Fachrichtungen Lehramt, Natur- und Ingenieurwissenschaften, Mathematik, Physik, und Informatik teil. Die offenen Aufgaben wurden mithilfe der qualitativen Inhaltsanalyse über ein Kategoriensystem mit zwei Ratern codiert. Zur Überprüfung der Modellstruktur wurden die Daten mit ein- und mehrdimensionalen Modellen der Item-Response-Theorie (IRT) analysiert. Die Dimensionalität des Kompetenztests wurde anhand von informationstheoretischen Maßen und der latenten Korrelationen zwischen den einzelnen Dimensionen untersucht und überprüft.

Die Ergebnisse sprechen für ein zweidimensionales Kompetenzstrukturmodell, bei dem die Komponenten *Informationsentnahmenähe Integration (MERI)* und *Konstruktionsnähe Integration (MERII)* die Daten am besten beschreiben. Latente Korrelationen deuten auf inhaltliche Zusammenhänge zwischen den Teilkompetenzen hin, lassen aber eine getrennte Betrachtung der Dimensionen MERI und MERII zu. Zwischen den Teilkompetenzen zeigen sich Unterschiede in den Kompetenzausprägungen. So ist für Studierende die *Konstruktionsnähe Integration (MERII)* schwieriger als die *Informationsentnahmenähe Integration (MERI)*. Regressionsanalysen bestätigten, dass die Komponenten MERI und MERII einen signifikanten Einfluss auf die Itemschwierigkeit haben, während für den *Repräsentationstyp* kein statistisch bedeutsamer Einfluss ermittelt werden konnte. Darüber hinaus hat der *Aufgabenkontext Genetik* einen statistisch bedeutsamen Einfluss auf die Itemschwierigkeit, was für die Kontexte Stoffwechsel und Ökologie nicht diagnostiziert werden konnte. Somit liegen erste Hinweise vor, dass der inhaltliche Kontext für die Schwierigkeit einer Aufgabe weniger entscheidend ist als die Anforderung an den Integrationsprozess beim Repräsentationswechsel selbst.

Damit knüpft die vorliegende Arbeit an bestehende Forschung zur Text-Bild-Integration an und leistet einen der ersten Beiträge zur fachdidaktischen Beschreibung und Untersuchung von biologiespezifischen Kompetenzen bei der Integration von multiplen externen Repräsentationen. Dadurch, dass die Aufgaben das Fähigkeitsspektrum der Oberstufe und Abiturwissen abfragen und bei Studienanfängerinnen und Studienanfängern eingesetzt wurden, liegen erste Hinweise vor, die zur Analyse von potentiellen Eingangsvoraussetzungen für ein naturwissenschaftliches Hochschulstudium herangezogen werden können. Damit verbunden sind Möglichkeiten der Kompetenzdiagnostik und Förderung, wenn es um repräsentationsspezifisches Wissen in den Naturwissenschaften geht.

I EINLEITUNG

Kommunikation im naturwissenschaftlichen Unterricht findet mit unterschiedlichen Formen der Informationsdarbietung statt, den sogenannten Repräsentationen (Schnotz, 2002).

“Reading, writing, and oral communication are critical literacy practices for participation in a global society” (Krajcik & Sutherland, 2010; S. 456).

In ihrem Beitrag in der international renommierten Zeitschrift *Science* betonen die Autoren über die Grenzen des Adressatenkreises in Lehrerbildung und Bildungsforschung hinaus, die Bedeutung von *Literacy* im naturwissenschaftlichen Unterricht, wenn es um die Teilhabe am gesellschaftlichen Diskurs und an der Meinungsbildung über Fortschritt und Forschung geht. In diesem Zusammenhang nimmt die Beherrschung naturwissenschaftlicher Fachsprache eine Schlüsselstellung ein (Lemke, 2004; Osborne, 2002; Yore, Bisanz, & Hand, 2003). Eine adäquate Fachsprache befähigt Schülerinnen und Schüler dazu, „sich über naturwissenschaftliche Themen zu informieren, naturwissenschaftliches Wissen zu erwerben [und] dieses Wissen zu kommunizieren (...)“ (Nitz et al., 2012, S. 118). Das Verstehen und Verarbeiten von Texten, Bildern und Symbolen gehören dabei zu grundlegenden Fähigkeiten in unserer Gesellschaft und sind Teil einer speziellen Kulturtechnik. Entsprechend kommt dem Biologieunterricht eine zentrale Rolle bei ihrer Vermittlung zu.

Gegenstand von Unterricht und Aufgabekultur sind häufig Text-Bild-Kombinationen, sogenannte *Multiple Externe Repräsentationen (MER)*. Dabei kommen verbale, visuelle und symbolische Repräsentationen allein oder in Kombination zum Einsatz. Aufgaben mit mehreren Darstellungsformen haben aufgrund ihrer Vielfältigkeit das Potenzial, biologische Inhalte mit einem lebensweltlichen Bezug darzustellen und so zur Ausgestaltung situierter Lernumgebungen beitragen zu können. Diese Darstellungen sind häufig nicht intuitiv verständlich und bereiten Lernenden verschiedener Altersstufen Schwierigkeiten (Ainsworth, 2006; Schnotz, 2002; Kozma & Russell, 1997; Anzai, 1991). Vor diesem Hintergrund erfährt die Integration von Text und visuellen Repräsentationen einen besonderen Stellenwert. Dabei wird explizit dazu aufgefordert: „teach students to use graphics and text to support meaning making“ (Krajcik & Sutherland, 2010, S. 457). Die Naturwissenschaft kann dementsprechend u.a. nur über ein „connecting (of) multiple representations“ (S. 457) erfolgen, da Lernende bei der Auseinandersetzung in einem Fachgebiet nicht nur eine Repräsentationsform nutzen müssen, sondern häufig Kombinationen aus verschiedenen Darstellungsformen. Die Lese- und Übersetzungsfähigkeit von MER sind für Lernende kognitiv herausfordernd, wobei gerade das Herstellen von Zusammenhängen zwischen den einzelnen Repräsentationen schwerfällt

(Seufert, 2003). Gleichzeitig sind diese Prozesse wichtige Voraussetzungen für das Verstehen und Kommunizieren naturwissenschaftlicher Strukturen, Konzepte und Prinzipien und entscheidend beim Problemlösen (Hettmannsperger, 2015; Prain & Tytler, 2012; Kozma & Russell, 2005; Kozma & Russell, 1997). Dabei stehen das „*cognitive linking*“ sowie das „*mapping between representations*“ (Ainsworth, 1999, S. 12) im Mittelpunkt nationaler und internationaler Studien und im Rahmen von Scientific Literacy. So ist es notwendig, dass Lernende verschiedene Repräsentationsformen aufeinander beziehen können. Dabei müssen Lernende zunächst jede Repräsentation für sich verstehen, verarbeiten und in ein mentales Modell integrieren (Mayer, 2014; Ainsworth, 2006; Schnotz & Bannert, 1999). Wurde der Umgang mit MER während der Schulzeit nicht ausreichend geübt, werden die Probleme in das Studium transferiert und können negative Folgen für das naturwissenschaftliche Konzeptverständnis haben (von Kotzebue & Nerdel, 2015). Übersetzungsleistungen zwischen Repräsentationen und der Umgang mit solchen MER werden damit zu einem konstituierenden Merkmal fachsprachlicher Kompetenz (Nitz et al. 2012) und eine zentrale Fähigkeit von Kommunikationsprozessen im naturwissenschaftlichen Unterricht. Entsprechend sollten Lernende zum Umgang mit MER und damit zur Interpretation, Konstruktion, Translation und Transformation durch den naturwissenschaftlichen Unterricht befähigt werden, da diese Aspekte zur Entwicklung von *representational competence* beitragen (Nitz et al., 2014; Kozma & Russell, 1997).

Bisherige Studien fokussieren beispielsweise ausschließlich einzelne *Externe Repräsentationen (ER)*, sind häufig repräsentationsspezifisch mit Beschränkung auf einen bestimmten Fachinhalt (z.B. Lachmayer, 2008; Nitz et al., 2012; von Kotzebue & Nerdel, 2015) oder beleuchten das Lernen mit Repräsentationen ohne die kognitiven Prozesse (z.B. Tsui & Treagust, 2013). Dadurch fehlen insbesondere Kompetenzmodelle zur Integration von MER, die der unterrichtlichen Verwendung von vielfältigen Darstellungsformen, z.B. in Arbeitsmaterialien und unterschiedenen Themenbereichen, Rechnung tragen. Hingegen konnten Schönborn & Anderson (2008) bereits für externe Repräsentationen innerhalb der Biochemie zeigen, dass eine erfolgreiche Interpretation von Repräsentation von verschiedenen Faktoren, wie z.B. dem Konzeptverständnis oder dem Verstehen des dargestellten Repräsentationstyps, abhängt. Da nicht alle Lerner in gleichem Umfang von MER profitieren (vgl. Ainsworth, 2006), ist es sinnvoll zu untersuchen, unter welchen Bedingungen ein Lernerfolg zu verzeichnen ist und welche Faktoren dabei einen Einfluss auf den Integrationsprozess haben. Ziepprecht (2016) erwähnt in diesem Zusammenhang, dass gerade die Charakteristika von Text-Bild-Kombinationen herangezogen werden sollten, wenn es um das Verstehen von fachspezifischen Besonderheiten beim Repräsentationsverstehen geht.

Vor diesem Hintergrund wurde ein fachdidaktisches Kompetenzstrukturmodell entwickelt und validiert, um Kompetenzen der integrativen Text-Bild-Kombination und deren Verarbeitung in unterschiedlichen biologischen Kontexten und verschiedenen Repräsentationstypen identifizieren zu können. Zu diesem Zweck werden in Kapitel II die theoretischen Grundlagen beschrieben und die theoretische Entwicklung des Kompetenzstrukturmodells zum Umgang mit multiplen externen Repräsentationen dargestellt. Die daraus hervorgehenden Forschungsfragen und Hypothesen werden in Kapitel III dargelegt. Der empirische Teil (Kapitel IV) berichtet über die Entwicklung der eingesetzten Testinstrumente und erläutert die Methodwahl zur Validierung des entwickelten Kompetenzstrukturmodells. Daran anschließend werden die Ergebnisse aus der empirischen Überprüfung des Kompetenzmodells zum Umgang mit MER vorgestellt. In Kapitel V werden die Befunde diskutiert. Die fachdidaktische Relevanz und der Ausblick auf die Unterrichtspraxis und weiterführende Forschung beenden die vorliegende Arbeit.

II THEORETISCHER HINTERGRUND

1 NATURWISSENSCHAFTLICHE KOMMUNIKATION

Naturwissenschaftliche Kommunikation heißt fachspezifische Kommunikation und steht im Zusammenhang mit dem Erlernen und der Verwendung von Fachsprache (Rincke, 2007). Grundlegend ist die Definition von kommunikativer Kompetenz nach Habermas (1971) sowie das Kommunizieren über fachbezogene Sachverhalte. Im Folgenden werden zentrale Aspekte naturwissenschaftlicher Kommunikation dargelegt. Ausgehend vom Konzept der *Scientific Literacy* werden naturwissenschaftliche Kommunikationsprozesse, die curriculare Relevanz vor dem Hintergrund der nationalen Bildungsstandards und die Bedeutung multipler externer Repräsentationen im naturwissenschaftlichen Unterricht vorgeschult.

1.1 SCIENTIFIC LITERACY

Scientific Literacy kann als ein grundlegendes Verständnis von Naturwissenschaften verstanden werden (vgl. Roberts, 2007; Norris & Phillips, 2003; DeBoer, 2000). Der Begriff bzw. das Konzept von Scientific Literacy wird in der Literatur hingegen nicht konsistent definiert und verwendet (Roberts, 2007). Gemeinsam ist allen Konzeptionen, dass naturwissenschaftliches Wissen ein Verständnis naturwissenschaftlicher Konzepte und Prozesse verlangt sowie Wissen über das Wesen und die Grenzen von Naturwissenschaft. Das Erkennen naturwissenschaftlicher Fragen, naturwissenschaftliches Arbeiten und das Ziehen von Schlussfolgerungen sind zentrale Forderungen von Scientific Literacy. Die Fähigkeit, wissenschaftlich zu denken und wissenschaftliche Erkenntnisse und Verfahren anzuwenden, ist bedeutsam für das bewusste Treffen von Entscheidungen sowie zur Reflexion und Erkenntnisgewinnung. Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) definiert Scientific Literacy als „an individual’s scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues“ (OECD, 2009, S. 14).

Die Definition zur naturwissenschaftlichen Grundbildungen umfasst drei zentrale Aspekte:

- I. Naturwissenschaftliche Prozesse beschreiben das Wissen, wissenschaftliche Phänomene zu beschreiben, zu erklären, zu verstehen und Erkenntnisse sowie Schlussfolgerungen zu ziehen. Damit erhält man das notwendige Wissen zur Lösung einer Aufgabe oder Problemstellung.
- II. Naturwissenschaftliche Konzepte bilden die Vernetzung und das Verständnis dieser wissenschaftlichen Phänomene und beziehen sich in PISA auf die naturwissenschaftlichen Schulfächer und somit auf anwendungsorientierte Aufgaben.
- III. Kontexte betreffen die Situationen in denen das Wissen und die Konzepte angewendet werden sollen und damit in den Aufgaben präsentiert werden.

Auf Basis dieser Definition und Kategorien werden die Testaufgaben für die PISA-Studien entwickelt. Naturwissenschaftliche Prozesse und Konzepte sind Bestandteil der Aufgabenkonstruktion und der Charakterisierung von Schülerleistungen, während Kontexte den fachlichen Hintergrund alltagsrelevant werden lassen (Baumert et al., 2001). Seit Einführung der PISA-Studien im Jahr 2000 und im Zuge der nationalen und internationalen Vergleichbarkeit von Schülerleistungen bilden diese drei Aspekte einen zentralen Fokus innerhalb der Kompetenzforschung und orientieren sich an dem übergeordneten Modell naturwissenschaftlicher Grundbildung (Baumert et al., 2001; Bybee, 1997). Unterschieden werden folgende Kategorien: 1. *Nominale Scientific Literacy*, d.h. die Kenntnis naturwissenschaftlicher Begriffe, 2. *Funktionale Scientific Literacy*, d.h. Faktenwissen und die korrekte Verwendung von Fachbegriffen als Grundlage für eine 3. *Konzeptionelle und Prozedurale Scientific Literacy*, die das Verständnis zentraler naturwissenschaftlicher Ideen, Fragen und Verfahren zusammenfasst und in eine 4. *Multidimensionale Scientific Literacy* führt, dem Verständnis der Besonderheiten naturwissenschaftlichen Denkens und die Integration in einen sozialen und kulturellen Zusammenhang (Bybee & McCrae, 2011; Schecker & Parchmann, 2006; Bybee, 1997). Naturwissenschaftliche Grundbildung wird damit zu einem Fundament für lebenslanges Lernen (Prenzel et al., 2001). Wissen muss also anschlussfähig, flexibel einsetzbar und anwendbar sein, d.h. es müssen Kompetenzen erworben werden, die das Aneignen neuer Wissensbestände und die Integration sowie die Auseinandersetzung mit bestehendem Wissen ermöglichen (Prenzel et al., 2001). Das naturwissenschaftliche Denken umfasst damit das Erkennen und Anwenden von naturwissenschaftlichen Fragestellungen, das Identifizieren von Daten sowie das Umgehen mit Evidenz, das Verständnis naturwissenschaftlicher Konzepte, aber auch die Fähigkeit naturwissenschaftliche Beschreibungen oder Argumente zu kommunizieren. Evidenzbasierte Schlussfolgerungen zu ziehen, gehört dadurch zur Basis naturwissen-

schaftlicher Kompetenz und umfasst neben Wissen auch Kommunikation (OECD, 2006; Norris & Phillips, 2003). Damit Lernende zu einem solchen naturwissenschaftlichen Verständnis kommen, also um „scientific literate“ zu sein, müssen sie fähig sein, naturwissenschaftliche Texte zu lesen und zu schreiben. Das Verstehen, Interpretieren, Analysieren und Kritisieren solcher Texte sind Bestandteil fachlicher Kommunikationsfähigkeit und damit Teil naturwissenschaftlicher Grundbildung (Norris & Phillips, 2003; Krajcik & Sutherland, 2010; Nitz, 2012). Während Bybee (2002) und die OECD (2006) die Konzeption von Scientific Literacy eher fachwissenschaftlich oder anwendungsbezogen definieren, liegt der Fokus bei Norris & Phillips (2003) auf der Literalität. In der Literatur wird zwischen grundlegender (*fundamental sense*) und abgeleiteter (*derived sense*) Scientific Literacy unterschieden (Norris & Phillips, 2003; Yore, Bisanz, & Hand, 2003; Yore, Primm, & Tuan, 2007). Unter einer grundlegenden Scientific Literacy versteht man das Lesen von Texten, während die abgeleitete Scientific Literacy das Wissen und Verständnis naturwissenschaftlicher Konzepte und Prozesse herausstellt und in der fachdidaktischen Literatur als *Fachsprache* bezeichnet wird. Chung et al. (2014) sehen diese Kommunikationsfähigkeiten als eine der wichtigsten Kompetenzen und Herausforderungen für das 21. Jahrhundert und betonen das Ziel „to enable students to articulate their ideas, thoughts, and feelings by using a variety of verbal and visual representations (e.g. words, images, gestures) and to deliver the key message taken from complex ideas“ (Chung et al., 2014, S. 2). Die Bedeutung von Scientific Literacy als multimodaler Diskurs betont auch Lemke (2004). Kommunikation entsteht durch Sprache und diese wird in Verbindung mit anderen semiotischen Darstellungen wie „visual images, diagrams, graphs, mathematical formulas, and the semiotics of artifacts, apparatus, and the meaningful activities of using them“ (Lemke, 2004, S. 1) ermöglicht. Dabei geht es um einen so genannten „visual turn“, d.h. um eine vielfältige Interpretation und Konstruktion verschiedener Darstellungsformen, die uns im täglichen Leben begegnen. Ob in der Zeitung, in Zeitschriften, auf Webseiten oder in Lehr- und Schulbüchern, wir werden stets dazu aufgefordert, diese Repräsentationen zu lesen, zu verstehen, zu interpretieren oder selbst zu konstruieren (Lemke, 2004). Entscheidend ist, dass es Aufgabe von Bildungsforschern ist, nicht nur das Lesen, Schreiben und Sprechen von Naturwissenschaft zu fördern, sondern vor allem zu erklären, warum Wissenschaftler welche Sprache zur Erklärung wissenschaftlicher Konzepte wann und wie einsetzen. Es geht darum, durch „verbal concepts, mathematical relationships, visual representations, and manual-technical operations“ (Lemke, 2004, S. 5) das Verstehen und Verständnis von Wissenschaftssprache und Fachsprache zu fördern. Die Sprache als Kommunikationsmittel ist also konstitutiv für (natur-)wissenschaftliches Handeln (Härtig et al, 2015). Damit wird Scientific Literacy über Bildung erworben (Crowell & Schunn, 2015) und

die formulierten Definitionen und Kompetenzen werden Gegenstand von Schule und Unterrichtsgestaltung. Eine Förderung adäquater Sprache muss damit Ziel für den naturwissenschaftlichen Unterricht sein. Orientierung bieten die gemeinsamen Standards der Länder (KMK, 2005a; KMK, 2005b), die das Rahmenkonzept Scientific Literacy verorten und wesentliche Aspekte von Fachsprache im Kompetenzbereich „Kommunikation“ aufgreifen.

1.2 KOMMUNIKATIVE KOMPETENZEN IN DEN BILDUNGSSTANDARDS

Im Jahr 2000 führte der PISA-Schock zu einer verstärkten Qualitätssteigerung und anschließender Outputorientierung in den darauffolgenden Jahren. Mit der Einführung verbindlicher, nationaler Bildungsstandards 2004 der „Ständigen Konferenz der Kulturminister“ wurden domänenspezifische Standards formuliert, die Schülerinnen und Schüler am Ende der Sekundarstufe I erreicht haben sollen. Zunächst für die Fächer Deutsch, Mathematik und Englisch folgte anschließend eine Erweiterung für die naturwissenschaftlichen Fächer Biologie, Chemie und Physik (KMK, 2005a; KMK, 2005b; KMK, 2005c). In vier Kompetenzbereichen, einer Inhaltsdimension (Fachwissen) und drei Handlungsdimensionen (Erkenntnisgewinnung, Kommunikation und Bewertung), werden domänenspezifische Kompetenzanforderungen formuliert und an Beispielaufgaben erläutert. Damit werden Lehr- und Lernziele beschrieben, „die einen Kompromiss zwischen der Orientierung an fachlicher Systematik, an funktionalen Anforderungen der Lebens- und Arbeitswelt und an den Lernvoraussetzungen und Entwicklungsbedürfnissen der Lernenden darstellen“ (Ferber, 2014, S.15). Die KMK betont dabei explizit, dass „inhalts- und handlungsbezogenen Kompetenzen (...) nur gemeinsam und in Kontexten erworben werden [können]“ (KMK, 2005a, S. 7). Kulgemeyer (2010) stellt heraus, dass die Standards nicht erkennen lassen, dass Kommunikation eine domänenspezifische Kompetenz darstellt, die für jedes Fach typische naturwissenschaftliche Aspekte zeigt und es sich damit um eine fachabhängige Schlüsselqualifikation handelt. Dennoch bieten die deutschen Bildungsstandards innerhalb des Kompetenzbereichs Kommunikation eine entscheidende Rahmenkonzeption zur Vermittlung und der angemessenen Verwendung von Fachsprache. Das Verstehen und der Umgang mit naturwissenschaftlicher Fachsprache ist eine Kommunikationsfähigkeit, die es Individuen ermöglicht, sich über naturwissenschaftliche Inhalte auszutauschen, neues Wissen anzueignen und dieses Wissen zu kommunizieren (Norris & Phillips, 2003; Yore, Bisanz, & Hand, 2003). Dadurch ist eine aktive Teilhabe am gesellschaftlichen Diskurs und innerhalb naturwissenschaftlicher Forschung, Technik und Entwicklung möglich. Kommunikationskompetenz mit dem „Ziel naturwissenschaftlicher Grundbildung (...), die Sprache (...) der Naturwissenschaften zu verstehen, [und] ihre Ergebnisse zu kommunizieren“ ermöglicht „das Erfassen und Vermitteln [z.B.] biologischer Sachverhalte“ (KMK,

2005a, S. 6). So ist die Kommunikation im Unterricht „direkter Lerngegenstand“ (KMK, 2005a, S. 11), „Mittel im Lernprozess“ (KMK, 2005a, S. 11) und grundlegend für das Erschließen naturwissenschaftlicher Inhalte aus fachsprachlichen Texten sowie Darstellungsformen im Sinne einer fachbezogenen Kommunikation. Kommunikationskompetenz ist eine der vier verbindlich vorgeschriebenen Kompetenzbereiche im Rahmen der KMK-Bildungsstandards und gilt für alle naturwissenschaftlichen Fächer (KMK, 2005a). Neben der Fachsprache ist auch die Alltagssprache der Schülerinnen und Schüler Bestandteil von Unterricht. Begriffe und Bezeichnungen werden häufig aus dem Alltag übernommen und erhalten eine neue Bedeutung (Stäudel, Franke-Braun, & Parchmann, 2008; Wellington & Osborne, 2001). Dabei sind Alltags- und Fachsprache immer auch von Schülervorstellungen und Vorwissen geprägt, was die Kommunikationskultur zwischen der Lehrkraft und den Lernenden zusätzlich erschweren kann (Kulgemeyer, 2010; Rincke, 2010). Die individuellen Alltagsvorstellungen der Lernenden sind Bestandteil des Fachunterrichts, gleichzeitig werden fachliche Konzepte und Fachsprache in die Alltagssprache zurückgetragen (KMK, 2005a). Diese Diskursfähigkeit ist somit expliziter Bestandteil der Bildungsstandards und Teil von Kommunikation. Aufgrund dieser Befunde muss naturwissenschaftlicher Unterricht diese unterschiedlichen Ebenen der Kommunikation berücksichtigen und einen adressatengerechten Weg der fachbezogenen Kommunikation im Rahmen der Literacy-Förderung gehen. Das wiederum heißt, die Unterrichtskommunikation zwischen Lehrkraft und Lernenden zu erhöhen, Fachsprache kritisch zu hinterfragen und ein domänenspezifisches Lesen von Sachtexten und Darstellungsformen zu fördern (Stäudel, Franke-Braun, & Parchmann, 2008). Dabei steht die Vielfalt des Angebots auch in keinem Widerspruch dazu, vorgegebene Standards für den mittleren Schulabschluss zu formulieren und gleichzeitig ein möglichst hohes Maß an fachbezogener Kommunikationskompetenz fördern und erreichen zu wollen. Die fachsprachlichen Kompetenzen können sich dabei an fachspezifischen Inhalten und Darstellungsformen entwickeln und werden in den nationalen Bildungsstandards berücksichtigt (Stäudel, Franke-Braun, & Parchmann, 2008). Der Kompetenzbereich Kommunikation steht für „Informationen sach- und fachbezogen erschließen und austauschen“ (KMK, 2005a, S. 7 ff.). In über zehn Standards wird festgehalten, über welche Fähigkeiten Lernende zum Erwerb von Kommunikationskompetenz verfügen sollen. Ziel ist es, dass Lernende „zu gesellschafts- oder alltagsrelevanten biologischen Themen [referieren]“ und dass diese Kommunikation und Argumentation „in verschiedenen Sozialformen“ (KMK, 2005a, S. 14 f.) abläuft. Dabei „werten [sie] Informationen zu biologischen Fragestellungen (...) zielgerichtet aus und verarbeiten diese (...) adressaten- und situationsgerecht“, um schließlich „biologische Phänomene [zu erklären]“ (KMK, 2005a, S. 15). Expliziten Bezug nehmen die Standards auf die Verwendung von Fachsprache und auf den Umgang

mit unterschiedlichen Darstellungsformen. So wird aufgeführt, dass Schülerinnen und Schüler in der Lage sein sollen, „den Bedeutungsgehalt von fachsprachlichen bzw. alltagssprachlichen Texten und von Bildern in strukturierter sprachlicher Darstellung [erklären und beschreiben können]“ und dabei „idealtypische Darstellungsformen, Schemazeichnungen, Diagramme und Symbolsprache auf komplexe Sachverhalte an[wenden]“ (KMK, 2005a, S. 15). Dies bezieht ebenfalls „Originale oder naturgetreue Abbildungen mit Zeichnungen oder idealtypischen Bilder“ (KMK, 2005a, S. 15) in die Ausbildung von Kommunikationskompetenz ein. In Anlehnung an Stäudel, Franke-Braun, & Parchmann (2008) können damit folgende Subbereiche zusammengefasst werden: Der Kompetenzbereich Kommunikation zielt auf eine aktive Förderung von Recherche-, Literatur- und Informationsarbeit ab, versucht den Lernenden Fähigkeiten und Fertigkeiten im Bereich eigenständiger Informationsentnahme und Konstruktion von Darstellungsformen bei adäquater Interpretation und Argumentation dieser zu vermitteln und gleichzeitig den Wechsel zwischen Alltags- und Fachsprache situations- und adressatengerecht zu unterstützen. Studien hingegen zeigen, dass gerade die Sprachaktivität im Unterricht seitens der Lernenden unterrepräsentiert ist, während die Lehrkraft das Unterrichtsgespräch dominiert (Seidel et al., 2006; Sumfleth & Pitton, 1998; Lemke, 1990). Durch diesen Mangel an fachbezogener Kommunikationskultur sowie am eigenständigen Konstruieren von Sprache als auch von Darstellungsformen im Unterricht kann davon ausgegangen werden, dass die Fachsprache nur hinreichend geübt oder aktiv erlernt wird und es damit zu fachsprachlichen Defiziten kommt (Lemke, 1990; Mortimer & Scott, 2000). Rincke (2011) stellt jedoch heraus, dass gerade der mündliche Austausch ein entscheidender Aspekt für den Erwerb von Fachsprache ist und empfiehlt eine Nutzung von fach- als auch alltags-sprachlicher Ausdrucksformen im Unterricht. Diese Tatsache betont die Notwendigkeit der Förderung eines aktiven Umgangs mit Fachsprache und Repräsentationen, indem naturwissenschaftliche Konzepte Gegenstand von Verbalisierungen im Unterricht werden und anhand typischer Darstellungsformen ein vernetztes fachbezogenes Wissen erlernt werden kann. Erst dadurch ist die Ausbildung fachsprachlicher Kompetenzen möglich, die wiederum Grundlage für eine fachliche Kommunikationsfähigkeit sind und Teil von Scientific Literacy (Norris & Phillips, 2003).

1.3 FACHSPRACHE

Fachsprache ist eine Schlüsselstellung von Scientific Literacy. Eine eindeutige Definition und Abgrenzung vom Begriff „Fachsprache“ gelingt aufgrund der vielfältigen Betrachtungsmöglichkeit bis dato nur sehr heterogen (Härtig, Bernholt, & Retelsdorf, 2015; Roelcke, 2014; Rincke, 2010). Sprache ist „an integral part of science and science literacy – language is a

means to doing science and to constructing science understandings (...) [and] is used to communicate about inquiries, procedures, and science understandings to other people so that they can make informed decisions and take informed actions" (Yore, Bisanz, & Hand, 2003, S. 691). Das Lesen und Schreiben naturwissenschaftlicher Texte sowie das Verstehen, Interpretieren und Kommunizieren naturwissenschaftlicher Erkenntnisse ermöglicht kritisches Denken und einen naturwissenschaftlichen Diskurs. Härtig, Bernholt, & Retelsdorf (2015) stellen heraus, dass Fächer aufgrund der sprachlichen Natur voneinander abzugrenzen sind und ein mentales Konstrukt darstellen. Dabei beziehen sich die Autoren auf Kalverkämper (1998) und darauf, dass durch fachliche Kommunikation erst Fachwissen dargestellt werden kann. In der Fachdidaktik ist der Begriff der „Fachsprache“ ein zentrales Element von naturwissenschaftlichem Unterricht und steht für eine fachbezogene Kommunikation (Härtig et al., 2015; KMK 2005a, b, c; Kalverkämper, 1998). Die Fachsprache kann aus einer deskriptiven respektive normativen Perspektive verstanden werden und beinhaltet ein bestimmtes Fachvokabular, syntaktische und stilistische Besonderheiten und eine Kommunikationssituation (Rincke, 2010). Roelcke (2010) gliedert Fachsprache in unterschiedliche Dimensionen. Die horizontalen (Fächer und Fachbereiche) und vertikalen (Abstraktionsebenen von Kommunikation) fachsprachlichen Varietäten sowie die sprachlichen Verwendungsarten (Fachtexte). Die horizontale Typologie bezieht sich dabei auf Sprache und Kommunikation innerhalb eines Faches, in der vertikalen Typologie stehen die Sprache und Kommunikation zwischen Experten und Novizen ein und desselben Faches bzw. zweier unterschiedlicher Fachbereiche. Die Typologie von Fachtextsorten bezieht schließlich symbolische und appellative Textfunktionen sowie eine konzeptionelle Textgestaltung ein. Bezugnehmend auf eine fachliche Kommunikation, der Förderung von mehr Kommunikationsfähigkeit und einer grundlegenden Scientific Literacy ist es notwendig, Fachsprache holistisch zu betrachten und Kommunikation als verbalsprachliche Äußerung und der Verwendung einer Vielzahl verschiedener Repräsentationen zu verstehen (Ainsworth, Prain, & Tytler, 2011; Yore & Hand, 2010; Prain & Waldrup, 2010; Lemke, 1990, 1998b, 2004). Die naturwissenschaftliche Fachsprache kann dabei als „a synergistic integration of words, diagrams, pictures, graphs, equations, tables, charts, and other forms of visual and mathematical expression“ (Lemke, 1998b, S. 3) beschrieben werden. Damit gehören neben verbalsprachlichen Repräsentationen auch gegenständliche, bildliche und symbolische Repräsentationen in eine holistische Betrachtung (Gilbert, 2007; Leisen, 2015; Lemke, 1998b). Einen besonderen Stellenwert erhält das Verbalisieren von Sprache, die neben (Fach-) Texten auch unterschiedliche Repräsentationen enthält. Leisen (2015) definiert sprachliche Standardsituationen im Unterricht „als sprachliche Lernumgebungen (...), in denen Sprachlernprozesse stattfinden“ (Leisen, 2015, S. 135). Anhand von vier sprachlichen Kompetenzbereichen fokussiert Leisen (2015) einerseits die Darstellung und

sprachliche Verhandlung von Wissen, den sprachlich begleitenden Wissenserwerb und den Ausbau von Text- und Sprachkompetenz. Dabei müssen die sprachlichen Lernprozesse materiell, d.h. in Form von Aufgabenstellungen und Materialien, sowie personell, d.h. durch Moderation der Lehrkraft, gesteuert werden. Kommunikationsfähigkeit im naturwissenschaftlichen Unterricht zu fördern, heißt damit den Erwerb von fachbezogener Sprachkompetenz zu ermöglichen (vgl. Muckenfuß, 1995). Dabei betonen die Autoren, dass der „Erwerb fachsprachlicher Kompetenzen (...) vom Erwerb komplexer – und lebenspraktisch nicht unmittelbar bedeutungsvoller – Theorieeinsichten nicht zu trennen (...)“ und „der Erwerb von Fachkompetenz unauflöslich mit ihrer Kommunizierbarkeit verknüpft wird“ (Muckenfuß, 1995, S. 246 und S. 266). Im Vergleich zur physikalischen Fachsprache, die, wie die Autoren herausstellen, nicht-kommunikativ ist (Muckenfuß, 1995), können für den Biologieunterricht und die biologische Fachsprache zwar zentrale Punkte „kommunikativer Kompetenz“ angenommen werden, gleichzeitig müssen aber auch die Besonderheiten der Domäne an sich berücksichtigt werden. Neben der Ausgangslage von Alltagssprache und Alltagsvorstellungen, dem Wissen oder Nicht-Wissen um Fachsprache und implizierter (Teil-) Theorien, gilt es neben einer Interpretation durch Konfrontation von Fachsprache und Alltagssprache, auch Sprachebenen dazwischen, wie die sog. Lernaltsprache bzw. Scientific Interlanguage, oder einen Metadiskurs, zuzulassen (Rincke, 2010; Lemke, 1990). Die Domäne der Biologie ist charakteristisch für Fachtexte, die morphologische und syntaktische Besonderheiten aufweisen sowie für eine Vielzahl an Darstellungsformen, die oftmals gleichzeitig präsentiert werden (Leisen, 2006). So sind z.B. Naturobjekte, Präparate, Zeichnungen und Schnittskizzen, Bilder, Anschauungs- und Funktionsmodelle, Säulen-, Linien- und Flussdiagramme, aber auch chemische Formeln sowie mathematische Gleichungen besonders charakteristisch für das Fach Biologie (Leisen 2015). Diese Darstellungsformen gehen einher mit einer Vielzahl von Fachbegriffen, d.h. das Fach Biologie zeichnet sich dadurch aus, dass es besonders terminusreich ist und gleichzeitig eine Vielfalt an externen Repräsentationen präsentiert wird (Enzinger, Nerdel, & Prechtel, 2012). Wirft man einen Blick in Fachbücher oder Schulbücher der Mittel- und Oberstufe, zeigt sich ein Sprachgebrauch, in denen Fachinhalte für unterschiedliche Adressaten aufbereitet sowie unterschiedliche Schwierigkeiten und Abstraktionsgrade dargestellt sind (Rincke, 2010). Im Jahr 1980 führen Bremer & Clemens (1980) für den Physikunterricht Befunde auf, die zeigen, dass ein Erlernen von Fachsprache dem Erlernen einer Fremdsprache gleicht. Die Menge und Vielfalt der Fachbegriffe in Schulbüchern und die Einführung neuer Begriffe im Unterricht erfordert von den Lernenden das Lernen einer Fremdsprachvokabel und gleichzeitiges Erlernen eines neuen Sachzusammenhangs (Wellington & Osborne, 2001). Das Verbalisieren innerhalb der biologischen Fachsprache ist damit eine besondere Herausforderung von naturwissenschaftlichem Unterricht und wird im Kompetenzbereich

Kommunikation der Bildungsstandards (KMK, 2005a) und den einheitlichen Prüfungsanforderungen für das Abitur (KMK, 2004) sowie im LehrplanPLUS für Gymnasien in Bayern besonders fokussiert, wenn es um „das Verständnis der biologischen Fachsprache und fachgemäßer Darstellungsformen [geht]“ (BStMUK/ISB, 2015).

1.4 EINTEILUNG VON REPRÄSENTATIONEN

Die Zunahme von bildlichen Repräsentationen gegenüber textuellen im Rahmen eines sogenannten „pictorial turn“, vom Text zum Bild, geht mit epistemischen Veränderungen in den Wissenschaften einher. So sind Bilder bereits historisch gesehen über die Kulturen hinweg stärker verbreitet als die Schrift (Deregowski, 1989). Repräsentationen sind zunächst unterschiedliche Formen der Informationsdarbietung (Schnotz, 2002) und können als Mittel und Zweck zur Verbalisierung von fachlichen Sachverhalten verstanden werden (Leisen, 2015). Unter Externen Repräsentationen (ER) versteht man Objekte der Außenwelt, die als Formen von „models, analogies, figures, diagrams, written or spoken text“ (Niebert & Gropengiesser, 2015, S. 5 f.) definiert werden und charakteristisch für die naturwissenschaftliche Fachsprache sind (Oliveira et al., 2015). Schnotz (2002) unterscheidet bei ER Texte und Bilder, die als deskriptionale und depiktionale Repräsentationen bezeichnet werden. Die Naturwissenschaften sind eine multicodele und multimodale Diskursgemeinschaft, die auf verschiedene Kommunikationsmittel Bezug nimmt. Multicodele Kommunikation basiert auf der Verwendung unterschiedlicher Symbolsysteme wie z.B. der Kombination von Text und Bild. Solche Text-Bild-Kombinationen, die aus mindestens zwei Repräsentationen bestehen, werden als multiple externe Repräsentationen (MER) definiert (Ainsworth, 2006). Unter einer multimodalen Kommunikation versteht man die Kommunikation mit mehreren Sinnesmodalitäten, z.B. auditiv und visuell wie es bei Videos der Fall ist (Weidenmann, 2002). Werden externe Repräsentationen in einem visuellen Modus dargestellt, spricht man von „visualization“ (Cheng & Gilbert, 2009). Repräsentationen kommen dabei in den unterschiedlichsten Formen vor, „including pictures (e.g., drawing, charts, graphs), written symbols (...), manipulative models, oral language (...), and real-world situations“ (Ryken, 2009, S. 348 nach Van de Walle, 2004). Beim Umgang mit unterschiedlichen Repräsentationsformen ist entscheidend, dass „Darstellungen (...) nicht von alleine [wirken], sie müssen gedeutet (gelesen, interpretiert) werden“ (Jörissen & Schmidt-Thieme, 2015, S. 385). Diese enorme Variabilität an unterschiedlichen Darstellungen, verbunden mit unterschiedlichen Abstraktionsgraden der enthaltenen Bilder, ist dabei eher die Regel als Ausnahme, wenn man in die Biologieschulbücher der Sekundarstufen blickt (Bergey et al., 2015; Roth & Pozzer-Ardenghi, 2013; Tsui & Treagust, 2013).

Daraus lässt sich ableiten, dass das Lernen und Lehren mit verschiedenen Darstellungsformen Voraussetzung für das Verständnis von Biologie ist (Bergey et al., 2015; Griffard, 2013). Die Visualisierung selbst ist zum einen Produkt (das Bild) und Prozess (der Akt der Visualisierung) zugleich (Ryken, 2009). Naturwissenschaftler verwenden Repräsentationen in ihrem täglichen Arbeitsalltag und nutzen diese im wissenschaftlichen Diskurs unter Fachkollegen bei der Präsentation wissenschaftlicher Ergebnisse und Erkenntnisse (Goldmann & Bisanz, 2002). Auf Basis des sprachlich beschriebenen oder bildlich gezeigten Sachverhalts, konstruiert das Individuum eine interne, mentale Repräsentation (Schnotz, 2002). Je nachdem, ob fachlich gesprochen oder geschrieben wird und in welchem Verhältnis die angebotenen Repräsentationen und die Adressaten stehen, sind Kommunikationssituationen unterschiedlich anspruchsvoll (Rincke, 2010). Hinzu kommt, dass sich in den Darstellungsformen die Domänenspezifität von Kommunikationskompetenz zeigt, d.h. Kommunikation im Fach eine fach- und adressatengerechte Kommunikation mit fachlichen graphischen Darstellungsformen ist (Kulgemeyer & Schecker, 2009). Anderson et al. (2013) formulierten in diesem Zusammenhang Fähigkeiten für die Biologie, die Schülerinnen und Schüler im Umgang mit Repräsentationen beherrschen sollen. Dabei geht es zunächst um die Dekodierung der symbolischen Sprache, aus der sich eine externe Repräsentation zusammensetzt. Darauf folgt eine Evaluation hinsichtlich Aussagekraft und Qualität. Externe Repräsentationen können zur Problemlösung herangezogen werden, müssen jedoch auch teilweise abgeändert werden, damit sie sich zur Interpretation eines biologischen Konzepts eignen. Des Weiteren sollen externe Repräsentationen von Lernenden selbst konstruiert werden, damit ein Konzept erklärt oder ein Problem gelöst werden kann. Aus diesem Grund sollte die Wahl der Darstellungsform immer vor dem Hintergrund der beabsichtigten Kommunikationsfunktion erfolgen (Schnotz, 2002). Trotzdem belegen Studien, dass Repräsentationen nicht nur über große Potenziale verfügen, sondern das Lernen von und mit (multiplen) externen Repräsentationen Schwierigkeiten und Probleme mit sich bringt. Vor diesem Hintergrund werden im Folgenden unterschiedliche Typen von Repräsentationen sowie ihre Funktionen, Potenziale und limitierenden Faktoren erläutert.

1.4.1 TEXTE

Texte werden als deskriptionale Repräsentationen bezeichnet und bestehen aus Symbolzeichen wie Lautzeichen und Schriftbilder. Sie haben keine Ähnlichkeit mit dem repräsentierenden Sachverhalt und zeigen eine arbiträre Struktur, da die Symbolzeichen ausschließlich durch eine Konvention, d.h. allgemein festgelegter Regeln, miteinander verknüpft sind (Schnotz & Bannert, 1999; Schnotz, 2002). Ballstaed et al. (1981) entwickelten ein Prozess-

modell zur semantischen Textverarbeitung. Darin werden Lesevariablen definiert, die Verständlichkeitsmessungen bezüglich der Lese-Text-Interaktion ermöglichen. Für Lernende bieten Texte einen entscheidenden Vorteil, sie sind eher konkret als abstrakt und können als Abfolge von Symbolzeichen einfach gelesen werden (Kintsch & van Dijk, 1978; Corradi, 2014). Eilam & Poyas (2010) berichten zudem, dass Lernende zur Lösung einer Aufgabe Informationen vor allem aus dem Text beziehen, auch wenn gleichzeitig Bilder angeboten werden. Mit Texten ist damit eine hohe Ausdrucksfähigkeit verbunden, da sich problemlos Negationen und Disjunktionen formulieren lassen (Schnotz, 2002). Fachtexte zeichnen sich durch einen unpersönlichen Stil aus und tragen Merkmale wie Funktionsverbgefüge, Gruppen von Nominalisierungen, Satzglieder sowie komplexe Attribute (Rincke, 2010). Die Wissenschaftssprache ist dabei besonders konventionalisiert und z.T. domänenspezifisch interpretierbar, wenn es um die Frage nach Alltags- und Wissenschaftssprache mit denselben Vokabeln und die enorme Fülle der Fachbegriffe geht (Kulgemeyer, 2009). Gerade Fachbegriffe können zu Verständnisschwierigkeiten bei Lernenden führen, was Analysen von Lehrmaterialien und Unterrichtssprache zeigen (Berck & Graf, 1992; Wellington & Osborne, 2001). Aus diesem Grund sind Prozesse des Textverstehens für die vorliegende Arbeit entscheidend, wenn entsprechend Texte mit Bildern verknüpft werden und im Speziellen eine Text-Bild-Integration erfolgen soll.

1.4.2 BILDER

Bereits Weidenmann (1994) zitierte Arnheim (1972) mit den Worten: „Wieviel wissen wir eigentlich darüber, was Kinder und überhaupt Lernende tatsächlich sehen, wenn ihnen eine Lehrbuchabbildung [...] vor die Augen komm(t)?“ (Arnheim, 1972, S. 290). Antwort finden wir bei der von den Bildautoren konzipierten adäquaten bildhaften Codierung, die für die Darstellung eines Sachverhalts getroffen und über Darstellungs- und Steuerungs-codes gebildet wird (Weidenmann, 1994). Bilder werden als depiktionale Repräsentationen bezeichnet (Schnotz, 1994). Sie sind aus ikonischen Zeichen zusammengesetzt und weisen dadurch eine wahrnehmbare Ähnlichkeit mit dem repräsentierenden Sachverhalt auf (Schnotz, 2005). Der Abstraktionsgrad kann sich zwischen einer depiktionalen Repräsentation und dem dargestellten Sachverhalt stark unterscheiden. Je nach Typ kann dies weniger abstrakt oder abstrakt mit starker Reduktion in den Details sein (Pozzer-Ardenghi & Roth, 2010). Zu dieser Auffassung kommen auch Treagust & Tsui (2013), indem sie Abbildungen im Rahmen eines Abstraktionskontinuums darstellen. Auch Schönborn & Anderson (2009) betonen ein solches Kontinuum, das von realistisch bis abstrakt reicht. Depiktionale Repräsentationen zeichnen sich allgemein durch eine rasche und effiziente Informationswahrnehmung, Informationsvollständigkeit sowie der Eigenschaft der Konsistenz aus, d.h. Depiktionen widersprechen sich

nicht (Schnotz, 2010). Mit Bildern werden in erster Linie Fotografien, Zeichnungen, Skizzen, Karten oder Ähnliches verbunden, man spricht von realistischen Bildern. In ihnen sind repräsentierte und repräsentierende Merkmale gleichwertig, d.h. durch die strukturelle Übereinstimmung existiert eine wahrnehmbare Ähnlichkeit (Schnotz, 2003). Ihre Dimensionen variieren dabei von realistisch zu schematisch und von konkret zu abstrakt. Logische Bilder wie Diagramme enthalten oft Inhalte, die nicht konkret visuell wie ein Gegenstand wahrnehmbar sind, aber eine zwischen Gegenstandsbereich und Bild strukturelle Analogie (z.B. abhängige und unabhängige Variable) aufweisen. Dies bedeutet, dass beliebige Merkmale eines Gegenstands durch räumliche Merkmale des Bildes dargestellt werden können (Schnotz, 2003). Obwohl Diagramme keine sichtbare Ähnlichkeit mit dem dargestellten Sachverhalt besitzen, stimmen sie dennoch in einer gewissen Hinsicht überein, indem die Merkmale innerhalb des Diagramms und innerhalb des abgebildeten Sachverhalts übereinstimmen (Schnotz, 2001). Logische Bilder zeigen eine arbiträre Struktur, die in hohem Maße konventionalisiert ist (Schnotz, 1994). Schemazeichnungen können als realistische Bilder mit hohem Grad der Realitätsnähe, d.h. wirklichkeitsgetreu vorliegen oder aber nur als schematisierte Umrisszeichnungen (Schnotz, 1994). Damit können Schemazeichnungen in den unterschiedlichsten Abstraktionsgraden bis hin zu modellhaften Darstellungen existieren. Die Übergänge zwischen logischem und realistischem Bild sind fließend und nicht immer eindeutig zuzuordnen. Eine exakte Definition und genaue Abgrenzung fehlen, weshalb Schemazeichnungen bisher kaum Gegenstand aktueller Forschung im Umgang mit multiplen externen Repräsentationen sind. Im Rahmen dieser Arbeit wird versucht Schemazeichnungen anhand existierender Merkmale genauer einzuordnen und voneinander abzugrenzen. Nach Sichtung zahlreicher Literaturquellen und der Analyse bestehender Einordnungen von Bildern, wird in der vorliegenden Arbeit folgende Einteilung getroffen (vgl. ABBILDUNG 1). Schemazeichnungen erhalten dabei eine eigene Untergruppe und werden entgegen erster Klassifikationen (vgl. Schnotz, 1994) nicht ausschließlich den realistischen Bildern zugeordnet. An dieser Stelle sei bereits auf das Kapitel 11 und 12 der vorliegenden Arbeit verwiesen. Im Rahmen der inhaltlichen Diskussion erfolgt eine gezielte Bildreflexion vor dem Hintergrund ausgewählter Beispiele und eingesetzter Aufgaben. Im Folgenden werden zunächst die unterschiedlichen Bildtypen, in der Reihenfolge zunehmender Abstraktion, von realistischen Bildern, Schemazeichnungen und logischen Bildern genauer vorgestellt und klassifiziert.

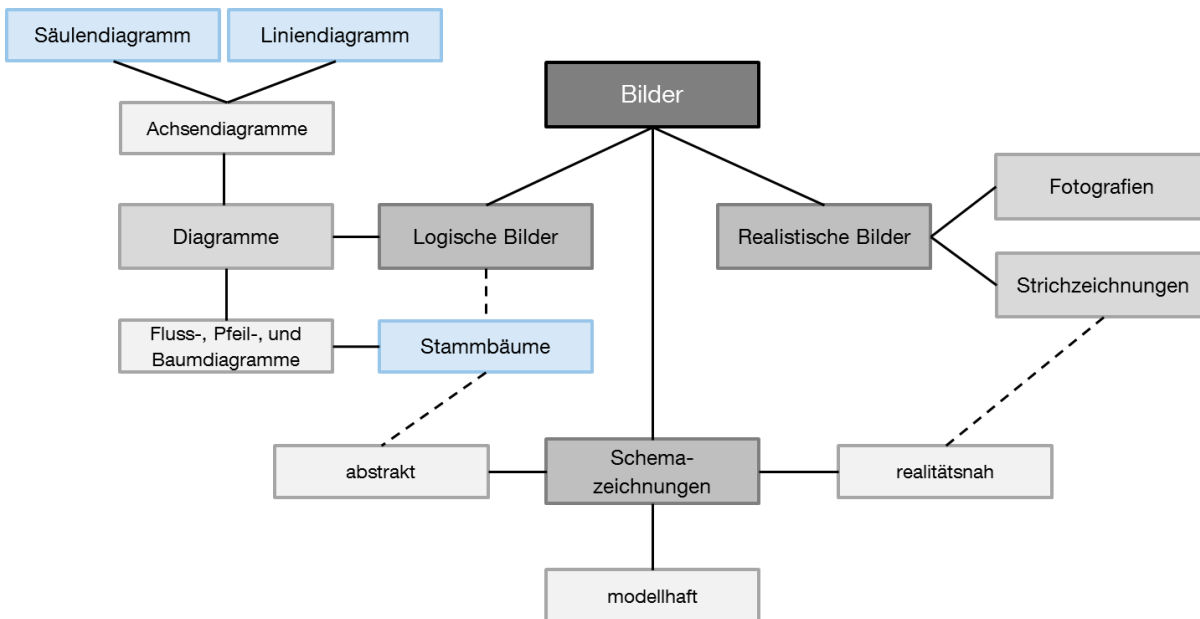


ABBILDUNG 1: EINTEILUNG VON BILDERN (ANGELEHNT AN SCHNOTZ, 1994)

1.4.2.1 REALISTISCHE BILDER

Unter realistischen Bildern werden Fotos, Strichzeichnungen, Piktogramme oder Landkarten eingeordnet, weil sie der dargestellten Realität ähneln und konkrete strukturelle Übereinstimmungen zwischen dargestelltem und repräsentierendem Gegenstand bestehen (Niegemann et al., 2008; Schnotz, 2002). Dadurch ermöglichen sie ein Erkennen des Abgebildeten anhand von Schemata der alltäglichen Wahrnehmung (Weidenmann, 1994; Schnotz et al., 2010). Sie visualisieren konkrete Realitätsausschnitte, enthalten jedoch bereits arbiträre Zeichen (Weidenmann, 1994). Realistische Bilder variieren in ihrem Abstraktionsgrad. Fotografien sind Abbildungen, die wie realistische, direkte Reproduktionen wirken. Strichzeichnungen repräsentieren Objekte, die durch graphische Mittel des Erstellers die wesentlichen Informationen darstellen (Weidenmann, 2002). Die Darstellungsobjekte sind zweidimensional, wobei Form, Kontur und Objektfarben variieren können. Dadurch werden auch Bewegungs- und Handlungsabläufe darstellbar. Beispiele sind biologische Prozesse wie Nährstoffkreisläufe oder der Ablauf der Proteinbiosynthese. Strichzeichnungen können durch die Nähe zum realen Darstellungsobjekt den Eindruck gewinnen, leicht verständlich zu sein. Häufig werden jedoch Sachverhalte repräsentiert, die nicht direkt erfahrbar sind und beispielsweise auf submikroskopischer Ebene ablaufen (Lynch, 1988, zitiert nach Brandstetter-Korinth, 2017). Dabei werden hochkonventionalisierte Zeichen und Symbole eingesetzt, die von Lernenden bewusst wahrgenommen und korrekt interpretiert werden müssen. Vor diesem Hintergrund werden derartige Strichzeichnungen als Schemazeichnungen verstanden, da sie in unterschiedlicher Darstellungsart abstrakte bis realitätsnahe Merkmale aufweisen können.

1.4.2.2 SCHEMAZEICHNUNGEN

Schemazeichnungen in Schul- und Lehrbüchern reichen von realitätsnahen bis hin zu stark reduzierten und abstrakten Darstellungen. Im internationalen Forschungsraum lassen sich biologische Schemazeichnungen den Fluss- und Pfeildiagrammen mit mehr oder weniger realistischen Anteilen in Form von Bildern zuordnen. Treagust & Tsui (2013) ordnen Schemazeichnungen auf einem Abstraktionskontinuum zu den Abbildungen mit zunehmender Abstraktion ein. Dabei enthalten Schemazeichnungen eine Mischung aus konkreten und abstrakten Elementen (Cheng & Gilbert, 2015), die wiederum „conventions and spatial isomorphism“ (Cheng & Gilbert, 2015, S. 138) wie z.B. Farben, Symbole oder Icons repräsentieren. Eilam & Poyas (2010) sprechen in diesem Zusammenhang davon, dass durch den Einsatz verschiedenster Mittel von Symbolen eine Vielzahl unterschiedlicher Schemazeichnungen in verschiedensten Beziehungen zueinander resultiert. Dabei haben sie immer Anteile von Text und Bild. Beim Lernen von biologischen Konzepten und zur Verbalisierung von Prozessen und Funktion werden häufig Schemazeichnungen verwendet, die bestimmte Strukturen darstellen (Cheng & Gilbert, 2015). Schemazeichnungen können dabei von semirealistisch bis logisch eingeordnet werden. Diese Definition erfordert eine präzise Eingrenzung im Rahmen der Einteilung von logischen und realistischen Bildern in der Abgrenzung von Schemazeichnungen und erklärt die hier vorgenommene Einteilung (vgl. ABBILDUNG 1). Biologische Schemazeichnungen sind besonders komplex, da sie häufig Prozesse dokumentieren und in verschiedenen Dimensionen und Organisationsebenen vorliegen sowie eine Vielzahl an Symbolen aufweisen (Griffard, 2013). Die Schemazeichnungen zum Wassertransport im Xylem oder zum Blutkreislauf werden international unter so genannten „complex process diagrams“ zusammengefasst (Griffard, 2013). Darunter fällt zunächst jede Form von graphischer Gestaltung, die vor allem in der Biologie bildliche Elemente enthält, die ikonisch oder semirealistisch sind und domänenspezifische Konventionen aufweisen (Griffard, 2013). Problematisch erscheint damit die Kombination aus Text, Symbolen (z.B. Pfeile) und Bildanteilen (von realistisch bis modellhaft) bei unterschiedlichem Vorliegen von nur einer Struktur oder eines gesamten Prozesses. So können die vorhandenen Konventionen innerhalb einer Schemazeichnung als dynamische, strukturelle Merkmale oder als Prozesse bzw. Verhaltensweisen dargestellt werden (Cheng & Gilbert, 2015). Brandstetter-Korinth (2017) untersuchte 608 Abbildungen in Biologieschulbüchern und konnte diese beiden Charakteristika, strukturell und prozesshaft, ebenfalls unterscheiden. Dabei werden biologische Inhalte einerseits strukturell als so genannte Objektdarstellungen mit visualisierender Text-Bild-Funktion und andererseits prozesshaft, d.h. als zusammengesetzte Strichzeichnungen, die eine organisierte Text-Bild-Funktion enthalten, dargestellt. Diese Komplexität führt zu einer großen Heterogenität und

macht eine exakte Einteilung von Schemazeichnungen schwierig. Im deutschsprachigen Raum ordnet Lachmayer (2008) Schemazeichnungen eindeutig den realistischen Bildern zu, während Oestermeier (2008) auf Autoren verweist, die Schemazeichnungen den Diagrammen zuordnen. Zu dieser Einteilung kommen aufgrund der eher arbiträren Strukturen mit geometrischen und topologischen Informationen in Schemazeichnung auch Ainsworth & Loizou (2003) sowie aufgrund der eher logischen Eigenschaften und dem symbolischen Gehalt auch Kosslyn (1989). Weidenmann (2002) weist ebenfalls darauf hin, dass Schemazeichnungen eine heterogene Gruppe sind und sowohl realistische bis hin zu ikonischen, domänenspezifisch auf Konventionen festgelegte Elemente enthalten können. Hinzu kommen die bereits beschriebenen Probleme der Terminologie im englischsprachigen Raum.

Im Rahmen dieser Arbeit werden Schemazeichnungen entsprechend ABBILDUNG 1 eingeteilt, voneinander abgegrenzt und zunächst als eigenständige Darstellungsform dargestellt. Zusätzlich werden die von Griffard (2013) formulierten Kriterien berücksichtigt, die für „complex process diagrams“ gelten. Zum einen zählen dazu Formen, die zur Repräsentation von biologischen Einheiten verwendet werden, d.h. illustrative, realistische sowie metaphorische Formen zur Repräsentation von z.B. Organismen, Zellen, Moleküle, Membranen. Daneben werden drei Dimensionen repräsentiert durch z.B. Schattierung, Schichtung oder Parallax. Die Zeit oder eine Sequenz wird mittels Pfeile dargestellt, die Anordnung erfolgt dabei in Leserichtung oder weist nummerierte Schritte auf. Verschiedene Organisationsebenen können durch Schachtelung von Vielfachen oder Übertreibung der Größe dargestellt werden. Gemäß den beschriebenen Kriterien können Stammbäume auch zu den abstrakten Schemazeichnungen gezählt werden, da sie bildliche Elemente enthalten, die ikonische und domänenspezifische Konventionen verlangen (Griffard, 2013). Geht man einen Schritt weiter, wird diese Abgrenzung erneut der bildlichen Definition nicht gerecht, indem die bildlichen Elemente eines Stammbaums als konventionalisierte Symbole verstanden werden müssen und die Leserichtung von verschiedenen Punkten aus begonnen werden kann. Aus diesem Grund verbleiben die Verbindungslinien in ABBILDUNG 1 in einer ersten Bildeinteilung gestrichelt und werden nicht als durchgezogen angenommen.

Cromley et al. (2010) definieren für das Verständnis von Schemazeichnungen bestimmte Voraussetzungen, ähnlich wie bei Diagrammen im Rahmen logischer Bilder. So sind für das Verstehen dieser Darstellungsformen inhaltliches und fachwissenschaftliches Wissen notwendig, aber auch das Wissen über Konventionen. Entsprechend erleichtert domänenspezifisches Vorwissen das Verständnis dieser Darstellungen. Dabei gelten Schemazeichnungen durch ihre farbliche Gestaltung interessanter als reiner Text, wodurch Lernende mehr Informationen für die aktive Informationsverarbeitung selektieren (Mayer, 1997; Ainsworth &

Loizou, 2003). Gleichzeitig wirkt die Darstellung der Information in Leserichtung als unterstützend, da der Lerner so eine lineare Wissensrepräsentation aufbauen kann (Stenning & Oberlander, 1995; Ainsworth & Loizou, 2003). Aus diesem Grund erfordern Schemazeichnungen eine besondere Aufmerksamkeit im Rahmen von Visual Literacy (Eilam, 2012). Sie sind immer dann lernförderlich, wenn der Lerner auf einen Sachverhalt direkt im Schema gelenkt und die Interpretation dadurch erleichtert wird (Griffard, 2013). Bergey et al. (2015) konnten in einer Interventionsstudie zeigen, dass Lehrkräfte das Lernen mit Schemazeichnungen durch Aufzeigen der relevanten schematischen Konventionen begünstigen können. Dieser konnte in besonders hohem Maße im Bereich der Biologie beobachtet werden. Schwierigkeiten beim Umgang mit Schemazeichnungen ergeben sich immer dann, wenn domänenspezifische Abbildungen verwendet werden und für den Lernenden unklar bleibt, welche Eigenschaften in diesem Schema zu erkennen sind (Cook, Wiebe, & Carter, 2008). Diese Schwierigkeiten stehen wiederum im Zusammenhang mit den Voraussetzungen. In einer Untersuchung von Cheng & Gilbert (2015) wurden Lernende der 10. Jahrgangsstufe einer Schule in Hong Kong mit Schemazeichnungen des menschlichen Blutkreislaufs konfrontiert. Die Autoren konnten durch die Antworten der Lernenden bestätigen, dass es vor allem die strukturellen Dimensionen in den Schemazeichnungen sind, auf die Bezug genommen wird. Ein übergeordnetes Verständnis über den eigentlichen Blutfluss, der über die Konventionen innerhalb der Schemazeichnung zu identifizieren ist, bleibt hingegen häufig unbeantwortet. Treagust & Tsui (2013) betonen ebenfalls, dass für das Verständnis von Schemazeichnungen multiple Repräsentationen koordiniert werden müssen, d.h. es muss eine Übersetzungsleistung zwischen den verbundenen einzelnen Repräsentationen stattfinden, die zusätzlich in verschiedenen Dimensionen vorliegen. Bergey et al. (2015) sprechen in diesem Zusammenhang erneut von einem „linking partially overlapping representations and sequencing the coordination process“ (Bergey et al., 2015, S. 2478), das von Ainsworth (2006) als eine Schlüsselqualifikation beim Lernen mit MER angesehen wird.

1.4.2.3 LOGISCHE BILDER

Per Definition dienen „logische Bilder [...] der Darstellung von Zusammenhängen zwischen qualitativen und quantitativen Merkmalen eines Sachverhalts, wobei es sich sowohl um wahrnehmbare als auch um nicht wahrnehmbare Merkmale handeln kann“ (Schnotz 1994, S. 97). Zu den logischen Bildern zählen vor allem Diagramme (im amerikanischen Englisch als *graphs* bezeichnet), die zur Veranschaulichung von ausschließlich quantitativen oder qualitativen und quantitativen Zusammenhängen dienen. Neben den Diagrammen werden auch *Charts*, die rein qualitativ sind, zu den logischen Bildern gezählt. Die Terminologie zeigt hier

keine einheitliche Struktur, so werden zu den Charts ebenfalls Tabellen gezählt, die im englischen Sprachraum auch als *diagrams* bezeichnet werden. Problematisch hierbei ist, dass zu *diagrams* auch schematische Abbildungen zählen. Diese werden laut Klassifikation nach Schnotz (1994) jedoch den realistischen Bildern zugeordnet. Aufgrund dieser Heterogenität sind Diagramme in der vorliegenden Arbeit logische Bilder, die sich auf die Darstellung von qualitativen und quantitativen sowie rein quantitativen Zusammenhängen beziehen. Schemazeichnungen sind dagegen eine eigene Gruppe von Bildern, die eine Zwischenstellung von logischen und realistischen Bildern einnehmen (vgl. ABBILDUNG 1).

Mit Diagrammen können abstrakte Sachverhalte dargestellt werden, die über realistische Bilder nicht vermittelt werden können. Damit Diagramme eine lernförderliche Wirkung zeigen können, müssen in ihnen Zusammenhänge und verbindende Elemente der vorliegenden Repräsentationen erkannt bzw. hergestellt werden können (Seufert, 2003; Kotzebue & Nerdel, 2012; Ainsworth, Bibby & Wood, 2002; Lachmayer, Nerdel, & Prechtel, 2007). Ein zentraler Unterschied bei der mentalen Verarbeitung zwischen Diagrammen und realistischen Bildern besteht in den kognitiven Schemata, die Einträge des Langzeitgedächtnisses repräsentieren. So sind kognitive Schemata, die alltägliches Wissen repräsentieren, für das Verstehen von logischen Bildern aufgrund der vorliegenden analogen Relation zwischen repräsentierendem und repräsentierten Sachverhalt wenig hilfreich, da kein Abgleich zwischen mentaler Repräsentation und den vorhandenen Schemata stattfinden kann. Diagramme sind daher nicht intuitiv verständlich, sondern stellen eine eigene Kulturtechnik dar (Dreyfus & Eisenberg, 1990; Schnotz, 2002). Bereits Leinhardt et al. (1990) berichten davon, dass sich der Prozess der Konstruktion von dem Prozess der Interpretation dahingehend unterscheidet, dass bei der Konstruktion immer neue Teile einer Darstellungsform generiert werden müssen, die noch nicht gegeben sind. Daraus resultiert auch, dass die Konstruktion je nach Anforderung und Komplexität unterschiedlich schwierig gestaltet werden kann (Leinhardt et al., 1990). Für die Informationsentnahme aus und die Konstruktion von Diagrammen sind damit spezifische Kompetenzen notwendig, die im Strukturmodell für Fähigkeiten beim Diagrammgebrauch von Lachmayer (2008) zusammengetragen werden (vgl. ABBILDUNG 2).

Informationsentnahme		Konstruktion		
Identifizierung	Erkennen der dargestellten Relation		Wahl des passenden Diagrammtyps	Aufbau des Rahmens
	Zuordnung der Variablen zu den Achsen		Zuordnung der Variablen zu ihren Achsen	
			Beschriftung der Achsen	
	Zuordnung der Datenreihen zu den Symbolen („Legende“)		Zeichnen einer Legende	
Beachten der Skalenreichweite		Zeichnen der Skalen		
Ablese	1. Ordnung	Ablese eines „Funktionswertes“	Eintragen der Punktwerte	Eintragen der Daten
	2. Ordnung	Vergleich zweier Werte oder Erkennen eines Trends (qualitativ/quantitativ)	Skizzierung einer Verbindungslinie zwischen Punkten* oder freie Skizzierung einer Trendlinie	
	3. Ordnung	Vergleich mehrerer Werte oder Vergleichen von Trends (qualitativ/quantitativ)	freie Skizzierung mehrerer Trends*	
		Extrapolieren/Vorhersagen		
Integration				

ABBILDUNG 2: STRUKTURMODELL FÜR FÄHIGKEITEN BEIM DIAGRAMMGEBRAUCH (LACHMAYER, NERDEL, & PRECHTL, 2007, S. 156, TAB. 1; LACHMAYER, 2008)

Im ersten Schritt der Informationsentnahme, erfolgt im Rahmen der Identifizierung, zunächst das Erkennen der dargestellten Relation sowie die Zuordnung der Variablen zu den Achsen. Dabei muss erkannt werden, welche Achse welche Variable repräsentiert. Anschließend kann die Datenreihe den entsprechenden Symbolen aus der Legende zugeordnet werden. Durch Beachten der Skalenreichweite und der Datenpunkte selbst kann die Relation genauer beschrieben werden. In einem zweiten Schritt gehören Ablesefähigkeiten zur Diagrammkompetenz. Diese werden in drei, in ihrer Schwierigkeit aufsteigenden Ordnungen, festgehalten. Dabei kann in der ersten Ordnung ein Funktionswert lediglich abgelesen werden, während die zweite Ordnung das Erkennen eines Trends beschreibt, indem zwei Werte verglichen werden. In einer höheren, der dritten Ordnung, wird der Vergleich von Trends beschrieben, die über Unterschiede zwischen mindestens zwei Werten identifiziert werden müssen. Die drei unterschiedlichen Ordnungen konnten empirisch nicht bestätigt werden, eine zunehmende Ableseordnung ist hingegen anzunehmen. Als weitere Fähigkeit im Diagrammgebrauch der Informationsentnahme steht das Extrapolieren, d.h. das Fortführen eines Trends auf Basis vorgegebener Datenpunkte. Die Kategorien der Informationsentnahme wurden von Lachmayer (2008) auf die Diagrammkonstruktion übertragen. Entsprechend erfolgt im ersten

Schritt der Aufbau des Rahmens. Hierzu müssen der passende Diagrammtyp gewählt, die Variablen korrekt den Achsen zugeordnet sowie die Achsen beschriftet werden. Des Weiteren ist eine Legende erforderlich und das Zeichnen der Skalen. Beim Eintrag der Daten in den Diagrammrahmen werden zunächst die Punktwerte eingetragen bevor eine Verbindungs- oder Trendlinie (im Falle von Liniendiagrammen) skizziert wird. Es können auch mehrere Trends frei skizziert werden. Untersuchungen zeigen, dass dabei häufig Fehler unterlaufen. Erfahrungsgemäß haben Lernende häufig Schwierigkeiten bei der Auswahl eines geeigneten Diagrammtyps. Sie können nicht entscheiden, ob ein qualitativer oder quantitativer Zusammenhang zwischen den Daten besteht und scheitern beim Eintragen von Punktwerten. Lernende kennen häufig nicht die Konvention, welcher Achse die abhängige und welcher Achse die unabhängige Variable zugeordnet wird. Eine weitere Schwierigkeit liegt in einer geeigneten Skalierung. Häufig ist die Skalierung unvollständig, unpassend gestreckt bzw. gestaucht (vgl. von Kotzebue, Gerstl, & Nerdel, 2015). Delgado & Lucero (2015) berichten von Schwierigkeiten beim Erstellen einer angemessenen Skalierung, wenn es um logarithmische Werte geht. Schwierigkeiten ergeben sich ebenfalls daraus, dass sich zwei oberflächlich ähnliche Diagramme in ihrer Aussagekraft wesentlich unterscheiden können. Während sowohl Linien- und Streudiagramme Zusammenhänge zwischen zwei Messgrößen zeigen, variieren sie dennoch, indem Liniendiagramme eine Entwicklung oder den Vergleich von Messwerten repräsentieren, während bei Streudiagrammen die Stärke der Übereinstimmung fokussiert wird. Lernende können Probleme in der Werteinterpretation haben, indem sie Werte im Diagramm falsch ablesen oder Intervalle als eigenständige Werte interpretieren. Schnotz (1994) berichtet, dass Novizen dazu tendieren lediglich einzelne Werte (häufig der maximale Wert) ablesen anstatt den Entwicklungstrend im dargestellten Diagramm zu erkennen.

Im Rahmen dieser Arbeit werden Diagramme in Form von Säulen- und Liniendiagramme verwendet (für weitere Achsendiagramme vgl. Schnotz, 1994). Grund für diese Auswahl ist, dass sich in den Schul- und Lehrbüchern zur Biologie größtenteils diese beiden Diagrammtypen wiederfinden und diese Diagramme damit einen ökologisch validen Untersuchungsgegenstand darstellen. Mit beiden Diagrammtypen werden quantitative Zusammenhänge visualisiert (Schnotz, 1994), wobei Säulendiagramme den Zusammenhang zwischen einer qualitativen (kategoriale Variable) und quantitativen Variablen darstellen. Liniendiagramme zeigen dagegen Zusammenhänge zwischen zwei quantitativen Merkmalen. Beide Diagrammtypen werden auch als Achsendiagramme bezeichnet, da auf den Skalen die jeweiligen Ausprägungen der Variablen angegeben sind. Eine Säule oder ein Datenpunkt lässt sich durch die relative Lage zu den Achsen eindeutig konkreten Werten zuordnen. Des Weiteren ist immer

mindestens eine Variable kontinuierlich (Lachmayer, Nerdel, & Prechtel, 2007). Sollen qualitative Zusammenhänge zwischen Variablen präsentiert werden, können Fluss-, Pfeil- und Baumdiagramme unterschieden werden (Schnotz, 1994). In ihnen sind einzelne Elemente wie z.B. Worte oder Symbole durch Linien und Pfeile miteinander verbunden, wobei Pfeile die Relationen zwischen den Elementen repräsentieren und gesondert interpretiert werden müssen. In diesem Zusammenhang könnte man annehmen, dass sich phylogenetische Stammbäume und Stammbäume mit Vererbungsmuster ebenfalls als solche Diagramme einordnen lassen. Betrachtet man hingegen derartige Pfeil- und Baumdiagramme genauer, lässt sich feststellen, dass hier kontextabhängige Bedingungen vorliegen, die sich auf verschiedene Inhaltsbereiche übertragen lassen (Lachmayer, 2008; Lachmayer, Nerdel, & Prechtel, 2007). Unter Flussdiagrammen werden Strukturen von Abläufen zusammengefasst, die vorgezeichnete Pfade aufweisen, die der Betrachter nachvollziehen kann (Weidenmann, 1994). Auch hier lassen sich Stammbäume mit Vererbungsmuster nur schwer verorten, zudem liegt bei Stammbäumen eine charakteristische, durch Konvention festgelegte Form, vor. Vor diesem Hintergrund müssten Stammbäume den Achsendiagrammen zugeordnet werden, was jedoch aufgrund der fehlenden Skalierung und Variablenausprägung ebenfalls nicht sinnvoll erscheint. Daraus schlussfolgernd fassen wir in unserer ersten Bildeinteilung Stammbäume als eine gesonderte Gruppe unter den logischen Bildern auf, da ein konventionalisierter Zusammenhang zwischen qualitativen und quantitativen Merkmalen eines Sachverhalts vorliegt (vgl. ABBILDUNG 1). Offen bleibt dabei die Möglichkeit einer gleichzeitigen oder gesonderten Einteilung in sogenannte abstrakte Schemazeichnungen.

1.4.3 TEXT-BILD-KOMBINATIONEN, MER

Multiple externe Repräsentationen (MER) werden ebenfalls als unterschiedliche Formen der Informationsdarbietung verstanden (Schnotz, 2002). Kombinationen aus mindestens zwei Repräsentationen bezeichnet Ainsworth (2006) als MER. Dabei werden sowohl Text-Text (Wörter und Symbole), Bild-Bild- als auch Text-Bild-Kombinationen in die Definition eingeschlossen. Gemäß den Gestaltungsfaktoren für multi-repräsentationale Systeme lässt sich die Form der Repräsentationen in Dimension, Abstraktion und Dynamik unterscheiden (Ainsworth, 2006). Wu & Puntambekar (2012) differenzieren MER als Kombinationen von mindestens zwei unterschiedlichen Typen von Repräsentationen.

Schönborn & Bögeholz (2009) berichten, dass für eine vollständige Interpretation einer MER jede einzelne darin enthaltene Repräsentation in ihren Konventionen zunächst für sich verstanden werden muss, damit ein „Verlinken“ zwischen einzelnen Repräsentationen ermöglicht werden kann. Entscheidend dabei sind vor allem die Prozesse aus Interpretation und

Integration, d.h. die Kommunikation über verschiedene externe Repräsentationen, in der „one form of representation to another“ (Mayer, 2002, S. 228) übersetzt und Informationen wechselseitig aufeinander bezogen werden müssen.

Das Lernen mit MER im Sinne von Text-Bild-Kombinationen und damit das Hinzufügen von Bildern, hat den Vorteil, dass Informationen besser behalten werden können. Lernen mit Text und Bildern ist damit dem Lernen mit nur reinem Text überlegen (Levie & Lentz, 1982). Man spricht auch vom Bildüberlegenheitseffekt (Weidenmann, 2002). Informationen, die über Bilder vermittelt werden, können effektiver als Langzeiteinträge gespeichert werden. Lernende können unter bestimmten Voraussetzungen mehr konzeptuelles Wissen aufbauen, wenn sie mit Bildern und Text anstatt mit reinem Text arbeiten (Mayer, 2001). Levin, Anglin, & Carney (1987) dokumentieren in ihren Ergebnissen einer durchgeführten Metaanalyse, dass das Lernen mit Text-Bild-Kombinationen, die darstellende, organisierende und interpretierende Bilder enthalten, einen Effekt auf das Behalten von Informationen hat und damit lernförderlich ist. Weitere Studien zeigen, dass MER im Vergleich zu einzelnen Repräsentationen das Lernen erleichtern können (Lem et al., 2015; Ainsworth, Bibby, & Wood, 2002; Schnotz & Bannert, 2003). Der Vorteil solcher bildhaften Darstellungen kann am kognitiven Aufwand zur Konstruktion mentaler Modelle liegen, der bei Bildern geringer als bei textbasierten Darstellungen ist (Schnotz & Bannert, 1999, s. Kapitel 2.4). Studien belegen außerdem, dass Lernende bei Text-Bild-Kombinationen höhergeordnete Strategien der Problemlösung (z.B. Zusammenfassen, Schlussfolgern) anwenden, als wenn ausschließlich Text gezeigt wird (Cromley et al., 2010; Griffard, 2013). Nach Saß (2012) scheint dies besonders bei komplexen Aufgaben zu stimmen. Die Studie konnte nachweisen, dass Bilder im Aufgabenstamm vor allem bei komplexen Aufgaben die Lösung positiv unterstützen. Hierzu passt, dass das reine Vorhandensein von Bildern zu keinem erhöhten Lerneffekt führt, während Text-Bild-Kombinationen mit konkreten Leseaufträgen zu einer Erhöhung führten. Dies entspricht der Annahme von Ainsworth & Loizou (2003), dass sich Lernen effektiver gestalten lässt, wenn in den Arbeitsmaterialien zwei unterschiedliche Modalitäten verwendet werden und diese durch das in Beziehung setzen ein höheres Verständnis ermöglichen (Cox, 1999).

Welche Funktionen MER haben und welche Potenziale sich beim Lernen mit MER entwickeln, werden von Ainsworth (1999; 2006) näher untersucht. Dabei beschreibt Ainsworth (1999; 2006) drei Hauptfunktionen von MER: *to complement*, *to constrain* und *to construct*. MER haben die Funktion sich „zu ergänzen“, d.h. durch den Gebrauch von zwei oder mehr Repräsentationen werden kognitive Prozesse unterstützt. Die übermittelte Information über MER kann dabei zunächst übereinstimmen oder sich unterscheiden. Dadurch vermitteln sich MER gegenseitig ergänzende Informationen, die einen Stimulationseffekt bewirken (Schnotz &

Bannert, 1999). Jede einzelne Repräsentation besteht aus einem eigenen spezifischen Informationsgehalt, der wiederum zu kognitiven Verarbeitungsprozessen führt. Für die Unterstützung im Lernprozess müssen MER jedoch den individuellen Bedürfnissen der Lernenden entsprechen und auf Anforderungen der Informationsentnahme abgestimmt sein. MER können das Lernen erleichtern, wenn der Interpretationsspielraum einer Repräsentation durch die Interpretation einer zweiten Repräsentation „eingeschränkt“ wird, d.h. durch eine zu hohe Ähnlichkeit oder durch inhärente Eigenschaften. Damit können Eigenschaften der einen Repräsentation genutzt werden, um die Interpretation einer weiteren Repräsentation anzuregen. Die dritte Funktion von MER ist der Aufbau eines vertieften Verständnisses durch gegenseitige Informationsanreicherung. Dies geschieht durch Abstraktionsprozesse, wodurch Lernende mentale Einheiten kreieren, die wiederum Ausgangspunkt für das Generieren neuer Konzepte auf einem höheren Organisationsgrad sind und an Kaput (1989) angelehnt sind: „The cognitive linking of representations creates a whole that is more than the sum of its parts (...)“ (Ainsworth, 1999, S. 141). Die gegenseitige Informationsanreicherung erfolgt über Konstruktions- und Ableseprozesse und kann das Wissen erweitern, indem Beziehungen zwischen MER erkannt werden und Lernende fähig sind, zwischen den Repräsentationen zu übersetzen. Diese Translationsprozesse können den Aufbau eines konzeptuellen Verständnisses fördern und bilden eine zentrale Komponente beim Umgang mit MER (Ainsworth, 1999; Nitz, Nerdel & Prechtel, 2012). Die Kombination unterschiedlicher Darstellungsformen wie im Rahmen von MER, kann bei Berücksichtigung dieser Funktionen zu einem Aufbau kohärenter Wissensstrukturen führen, wenn die Zusammenhänge erkannt werden (Seufert, 2003). Lernende können dadurch zwischen unterschiedlichen Repräsentationen wechseln, d.h. sie übersetzen, identifizieren verbindende Elemente und können die dargestellten Informationen integrieren sowie über Form und Funktion diskutieren (Ainsworth, 1999; 2006; Seufert, 2003). Die beschriebenen Funktionen hängen dabei vom Wissensstand und den Zielen der Lernenden ab, werden sie jedoch aktiv genutzt, können MER Wissen aufbauen, anstatt es einfach nur widerzugeben (Ainsworth, 2006; Yore & Hand, 2010). Cox (1999) nennt in diesem Zusammenhang drei Einflussfaktoren auf das Lernen mit Repräsentationen. Erstens die Eigenschaften der Repräsentationen, zweitens die Aufgabenanforderung und drittens innerpersönliche Merkmale wie Vorwissen und Art der Wissensaneignung.

Des Weiteren können MER einen Wissenstransfer im Rahmen des Kompetenzbereichs Fachwissens ermöglichen, indem Translationsprozesse beim Umgang mit MER in den Kompetenzbereich Kommunikation eingeordnet und mit den fachlichen Inhalten in Beziehung gebracht werden (Schönborn & Bögeholz, 2009). Die Autoren verweisen dabei auf die Forderung der Kultusministerkonferenz, dass Lernende im Rahmen der Kompetenzbereiche über

ein so genanntes „interconnected content knowledge“ (Schönborn & Bögeholz, 2009, S. 934) verfügen sollen (KMK, 2005; Harms et al., 2004). In der Studie von Kozma & Russell (1997) wurde die Rolle der fachlichen Expertise beim Verstehen von Text-Bild-Kombinationen untersucht. Ziel der experimentellen Studie war ein Experten-Novizen-Vergleich bei der Gruppierung und der Translation chemischer Repräsentationen. Im ersten Experiment wurden die Probanden beim Lauten Denken aufgefordert, unterschiedliche Darstellungsformen wie Grafiken, Videos und Reaktionsgleichungen zu ordnen. Es zeigte sich, dass Experten die Repräsentationen in größeren Gruppen zusammenfassen als die Novizen. Beide Gruppen argumentieren die Ordnungsstruktur inhaltlich oder aufgrund von Oberflächenmerkmalen. In einem zweiten Experiment wurden die Probanden aufgefordert, gegebene Repräsentationen in einen anderen Repräsentationstyp umzuwandeln oder aus mehreren Repräsentationen die Äquivalente auszuwählen (Kozma & Russell, 1997). Signifikante Unterschiede konnten beim Translationsprozess, d.h. der aktiven Übersetzung zwischen den Repräsentationen, nachgewiesen werden. Es zeigte sich, dass die Experten vor allem bei der Verbalisierung einer Repräsentation deutlich besser abschneiden als die Novizen.

Auch Schnotz (2002) sowie Schnotz & Bannert (1999) konnten zeigen, dass vor allem Lernende mit geringem Vorwissen von MER profitieren, da der kognitive Aufwand für die Erstellung eines mentalen Modells beim Lesen und Verstehen eines Textes höher ist, als beim Verstehen eines Bildes, das die gleiche Information vermittelt. In der Studie von Cook, Wiebe & Carter (2008) waren die Probanden mit höherem Vorwissen in der Lage inhaltliche Bezüge zwischen einer großen Anzahl von Repräsentationen herzustellen und konnten dadurch das konzeptuelle Verständnis der Repräsentationen untereinander fokussieren. Bildliche Darstellungsformen können so das Problemlösen erleichtern, indem sie Suchprozesse und die Arbeitsgedächtniskapazität verringern. Vor allem anspruchsvolle Bilder führen zu einer wechselseitigen Stimulation zwischen Text und Bild, indem Kohärenzlücken entweder im Text oder im Bild geschlossen werden (Schnotz & Bannert, 1999). Verbale Repräsentationen verlangen hingegen mehr aktive Suchprozesse sowie die Inferenz als bildliche Repräsentation (Cox, 1999). Mayer (1997) fügt hinzu, dass Bilder immer dann erklärende Funktion haben, wenn sich die Information im Text und im Bild aufeinander beziehen sowie in räumlicher und zeitlicher Nähe präsentiert werden (Schnotz & Bannert, 1999).

In der Studie von Schönborn & Bögeholz (2009) wurden zehn Experten darüber gefragt, welche Rolle das Translatieren und Transformieren biologischer Informationen spielt und welches Wissen hierzu von den Lernenden benötigt wird. Translation wurde in Anlehnung an Ainsworth (1999; 2006), Kozma & Russell (1997) und Prain & Waldrip (2006) definiert als „The ability to move across, interpret, and, in a multi-directional manner, link between ERs that

represent an underlying biological concept, principle or process at a particular level of biological organization“ (Schönborn & Bögeholz, 2009, S. 937). Unter Transformation versteht man “The ability to transfer knowledge from one situation to another (...). ‘Horizontal’ transfer is the successful application of biological knowledge from one context to another at the same level of biological organization. ‘Vertical’ transfer is the successful application of biological knowledge from one level of biological organization to another” (Schönborn & Bögeholz, 2009, S. 923). Die Experten sind der Meinung, dass das Aneignen von Wissen und der Wissensaufbau selbst ein vielseitiger und komplexer Prozess ist, der sowohl Kenntnisse auf horizontaler und vertikaler Ebene verlangt. Zusätzlich argumentieren die Experten, dass das Translatieren die Verarbeitung und anschließende Interpretation der äußeren Merkmale der dargestellten Repräsentation erfordert. Diese Ansicht wird für die Biologie weiter ausformuliert, indem Translationsprozesse als ein „moving across more than one ER that conveys the same [and different] biological idea[s]“ (Schönborn & Bögeholz, 2009, S. 943) dargestellt werden. Als Voraussetzung fassen die Autoren folgende vier Wissensarten zusammen: biologische Fachbegriffe, biologische Konzepte, das zugrundeliegende biologische Prinzip und Basiskonzepte (Schönborn & Bögeholz, 2009).

Aus diesem Grund sind die unterschiedlichen Informationsdarbietungen aufeinander zu beziehen, indem gemeinsame Strukturmerkmale identifiziert oder die unterschiedlichen Sachverhalte, die in den Repräsentationen verdeutlicht werden, sinnhaft miteinander verknüpft werden, um ein kohärentes mentales Modell zu erzeugen. Dabei gilt es zu beachten, dass explizite Unterstützungsmaßnahmen für das Verständnis von Bildern angeboten werden, um Schwierigkeiten entgegenzuwirken und eine lernförderliche Wirkung zu erzielen (Seufert, 2003). Werden diese Aspekte berücksichtigt können MER auch situierte Lernumgebungen ausgestalten, indem fachliche Inhalte mit großer Nähe zur Lebenswelt dargestellt werden. MER ermöglichen einen vertikalen und horizontalen Perspektivenwechsel, d.h. einen Wechsel zwischen verschiedenen Themen, aber auch zwischen den verschiedenen Systemebenen der Biologie (Tsui & Treagust, 2013). Repräsentation können damit helfen die Organisationsebenen des Lebendigen innerhalb der Biologie adäquat darzustellen (Treagust & Tsui, 2013). Für eine solche Darstellung sind MER unerlässlich. So können die anatomische und mikroskopische Organisationsebene mithilfe von Mikroskopie-Fotografien oder Zeichnungen dargestellt werden. Für molekulare Prozesse können Kreisläufe anhand von Schemazeichnungen dargestellt sowie Animationen oder Simulationen herangezogen werden. Durch textbasierte Elemente können einerseits wichtige Kernaussagen zusammengefasst oder weitere Informationen im Rahmen der Aufgabenstellung bereitgestellt werden.

Daneben zeigen zahlreiche Studien, dass das Lernen mit MER nicht nur positive Faktoren hat, sondern auch mit Schwierigkeiten verbunden ist. So haben Lernende häufig Probleme, wenn es darum geht, Informationen aus Text und Bild zu kombinieren und zu integrieren (Brünken, Seufert, & Zander, 2005; Levie & Lentz, 1982; Weidenmann, 1994).

In ihrer Studie untersuchten Watzka & Girwidz (2015) 177 Realschülerinnen und –schüler der neunten Jahrgangsstufe zum Einfluss der Kontextorientierung und des Präsentationsmodus von Aufgaben auf den Wissenserwerb und die Transferleistung von physikalischen Inhalten. Dazu wurden die Probanden in zwei Gruppen eingeteilt, die entweder einen stark oder einen weniger authentischen Kontext erhielten. Zusätzlich wurden die Gruppen nach textbasierten Aufgaben, bildbasierten Aufgaben und text-bildbasierten Aufgaben unterschieden. Die Analysen konnten keine Effekte der Kontextorientierung auf den Wissenserwerb nachweisen. Hingegen zeigen die Ergebnisse, dass die kognitiven Lernaktivitäten bei den Lernenden mit bildbasierten Aufgaben am stärksten ausgeprägt waren (Watzka & Girwidz, 2015). Diese Gruppe schnitt auch am besten in den Übungsaufgaben ab. Dies widerspricht allerdings den Annahmen von Weidenmann (2002), Schnotz (2005) und Mayer (2009), dass durch die Verknüpfung zwischen verbalen und visuellen Repräsentationen im Arbeitsgedächtnis eine lernförderliche Wirkung eintritt. Engelkamp (1990) begründet in seiner Gedächtnistheorie, dass bildhaft präsentierte Informationen direkter verarbeitet werden können als textbasierte. Dies stimmt mit dem Modell von Schnotz (2005) überein, dass die Verarbeitung von reinen bildbasierten Aufgaben direkt zum Abruf eines mentalen Modells führt, während bei der Verarbeitung von Texten zunächst ein propositionales Netz notwendig ist. Schnotz et al. (2011) ergänzen in diesem Zusammenhang, dass die Integration von verbaler und textlicher Information ein Übersetzen zwischen den entsprechenden Elementen in Text und Bild erfordert und erst durch dieses „mapping“ zunächst eine (wahrnehmbare) Oberflächenebene und eine anschließend (semantische) Tiefenebene gebildet werden kann. Damit kann nicht bestätigt werden, dass Lernende die höchste kognitive Aktivierung vor allem durch Text-Bild-Kombinationen empfinden (Watzka & Girwidz, 2015). Auch Rubitzko (2006) berichtet in der durchgeführten Interventionsstudie mit Schülerinnen und Schülern der 9. Jahrgangsstufe (N=116), dass bei einer Informationsentnahme aus mehreren Repräsentationen die Aufgaben weniger gut gelöst wurden als bei Aufgaben mit nur einer Repräsentation. Dafür sind komplexe Verknüpfungsprozesse erforderlich.

Rau, Alevén, & Rummel (2015) untersuchten in ihrer Studie das Lernen mit multiplen graphischen Repräsentationen (MGR). Dabei berücksichtigen sie neben der Cognitive Theory of Multimedia Learning (CTML, s. Kapitel 2.3) auch die Theorie des Integrativen Bildverständ-

nisses (ITPC, s. Kapitel 2.4) nach Schnotz & Bannert (2003). Die Autoren konnten die Annahmen beider Theorien bestätigen, d.h. dass der Lernerfolg mit Repräsentationen davon abhängt, inwieweit die Informationen in ein kohärentes mentales Modell integriert werden können (vgl. Mayer, 2005; Schnotz, 2005). Hinsichtlich eines besseren Lernverständnisses von multiplen im Vergleich zu einzelnen graphischen Repräsentationen konnte die Studie keinen signifikanten Vorteil nachweisen. Damit bleibt festzuhalten, dass zum Aufbau wechselseitiger Beziehungen zwischen deskriptionalem und depiktionalem Zweig der Text-Bild-Integration Konstruktions- und Ableseprozesse entscheidend sind. Ainsworth (2006) stellt heraus, dass der Schwierigkeitsgrad bei der Integration von MER mit zunehmenden Unterschieden zwischen den Darstellungsformen und den enthaltenen Strukturmerkmalen stetig ansteigt. Der Umgang mit einzelnen Repräsentationen und die zur Kohärenz führende Integrationsleistung muss also zunächst eingeübt werden. So ist eine erfolgreiche Text-Bild-Integration nicht nur von der Text- oder Bildkomplexität, vom Vorwissen oder dem kompetenten Leseverständnis der Repräsentationen abhängig, einen entscheidenden Beitrag kommt der Bereitschaft zu einer tiefen semantischen Verarbeitung zu (Schnotz et al., 2011, s. Kapitel 2.4). Werden Lernmaterialien ungeeignet gestaltet, kann ein Split-Attention-Effekt entstehen, der aufgrund einer kognitiven Überlastung zu einer reduzierten Lernleistung führt (Tabbers, 2002). Schnotz & Bannert (1999) stellen in diesem Zusammenhang fest, dass die Gestaltung von Lernmaterial mit nicht aufgabenäquivalenten Bildern zur Interferenz zwischen Bildstruktur und Modellstruktur führt. Voraussetzung für das Lernen mit MER ist damit, dass Lernende zunächst jede Repräsentation für sich verstehen (Ainsworth, 1999). Seufert (2003) spricht hier von „intra- und inter-representational coherence formation“ (Seufert, 2003, S. 228), d.h. ein Verständnis „within“ und „between“ der dargestellten Repräsentationen. Anschließend müssen sie in der Lage sein, die Beziehungen zwischen den verschiedenen Repräsentationen zu identifizieren, zu verknüpfen und die Reproduktion der ausgemachten Relationen zu verstehen (Ainsworth, 1999; Seufert, 2003). Allerdings zeigen sich hier immer wieder Schwierigkeiten (vgl. Bodemer, 2004; Bodemer & Faust, 2006; Baumert, Bos & Lehmann, 2000; Kozma & Russell, 1997; Anzai, 1991). Eine unzureichende Übersetzung zwischen verschiedenen Darstellungsformen verhindert ein tieferes Verständnis des dargestellten Sachverhalts. Gerade Novizen in einem bestimmten Fachbereich haben hierbei Probleme (Schwonke, Berthold, & Renkl, 2009).

Neben Problemen innerhalb der kognitiven Informationsverarbeitung sind weitere Einflussfaktoren auf das Lernen und Verstehen von MER entscheidend. Cox (1999) konnte zeigen, dass der Lernerfolg beim Lösen von Aufgaben mit Text-Bild-Kombinationen von den individuellen Lernvoraussetzungen variiert. Vor allem Versuchspersonen mit geringem Vorwissen konnten von MER profitieren, während bei Probanden mit fachlichen Vorkenntnissen kein

signifikanter Unterschied zu verzeichnen war (Mayer, 1997). Grund hierfür ist, dass Lernende mit fachlichem Vorwissen visuelle Vorstellungen und mentale Modelle auch rein durch Texte bilden können. Novizen bilden zwar ebenfalls mentale Modelle der dargebotenen Information, sie unterscheiden sich jedoch dadurch, dass sie häufig nur Oberflächenmerkmale erkennen und entscheidende Strukturen nicht (Seufert, 2003). Kozma & Russell (1997) berichten ebenfalls, dass Novizen unterschiedliche Repräsentationen nur über deren Oberflächenmerkmale aufeinander beziehen, Experten hingegen Bezüge über dahinterliegende Konzepte herstellen können. Stern, Aprea, & Ebner (2003) verweisen dabei auf das mangelnde Fachwissen, das es den Lernenden nicht erlaubt erfolgreich zwischen den Darstellungsformen zu übersetzen. Ainsworth (2006) weist auf das Problem der mangelnden Vertrautheit der Repräsentation und der fachlichen Inhalte hin und betont, dass dadurch gerade Tiefenstrukturen von Repräsentationen nicht verstanden werden können. Nitz et al. (2014) berichten, dass zwischen der inhaltlichen Wissensdomäne und der Repräsentationskompetenz ein korrelativer Zusammenhang existiert. Ein weiteres Hindernis für Lernende mit geringem Vorwissen ist die fehlende Bildung von Chunks im Arbeitsgedächtnis, sodass nur geringe Mengen von Wissens-elementen ins Arbeitsgedächtnis transportiert werden können, wodurch die Kapazitätsleistung sinkt und der Intrinsic Cognitive Load steigt (Seufert, 2003; s. Kapitel 2.3). Im Gegensatz dazu können Lernende mit größerem Vorwissen vor das Problem gestellt werden, dass sie Repräsentationen lediglich so interpretieren, wie es zu ihrem Vorwissen passt, d.h. sie verknüpfen diese mit bereits bekannten Strukturen und kognitiven Schemata, wodurch Missverständnisse generiert werden können (Lachmayer, Nerdel, & Prechtel, 2007). Ainsworth (2006) berichtet in diesem Zusammenhang, dass Lernende häufig die spezifischen Strukturmerkmale einer bestimmten Repräsentationsform nicht erkennen, jedoch Merkmale einer bestimmten Repräsentation verwenden, um eine völlig andere Repräsentation zu beschreiben. In ihren mit Eyetracker durchgeführten Studien belegen Schwonke, Berthold, & Renkl (2009) die vermittelnde Funktion von Bildern bei Text-Bild-Kombinationen und bei der aktiven Übersetzungsleistung. Die Autoren kommen ebenfalls zu dem Schluss, dass das Bewusstmachen der Funktionen der verschiedenen Darstellungsformen sich positiv auf das Lernen auswirkt, ohne dabei mehr Zeit zu verschwenden. Dieses Bewusstmachen ist damit ein essentielles didaktisches Element, damit Darstellungen effektiver, effizienter und bewusster wahrgenommen werden können (Schwonke, Berthold, & Renkl, 2008). In diesem Zusammenhang stehen auch verschiedenen Studien, die den Einfluss verschiedener Lernstrategien auf das Bearbeiten von Text-Bild-Kombinationen untersuchen (vgl. Ainsworth & Loizou, 2003; Schlag & Plötzner, 2011; Ziepprecht, 2016).

1.5 BIOLOGIESPEZIFISCHE REPRÄSENTATIONEN UND KONTEXT

Der Biologieunterricht zeigt eine enorme Vielfalt von biologiespezifischen Repräsentationen. Entscheidend sind dabei kontextspezifische, kommunikative Kompetenzen in Lern- und Leistungssituationen (Klieme & Leutner, 2006). Kompetenz wird damit als erlernbare Fähigkeit in fachlichen Zusammenhängen vermittelbar (Grube, 2010). Demzufolge werden von Lernenden Anforderungen in unterschiedlichen Kontexten gefordert, worunter die Teildisziplinen verschiedener Schulfächer verstanden werden. Ziel dieser Kontextorientierung ist die sachlogische Erschließung eines Teilgebiets, eine thematische Aspektierung eines fachlichen Teilgebiets innerhalb einer Wissenschaftsdomäne (Muckenfuß, 2004) und eine damit verbundene Vereinfachung, Verdeutlichung sowie Anwendbarkeit von Fachinhalten (Gilbert, 2006).

Nitz, Nerdel, & Prechtel, 2012 fassen beispielsweise verbale, bildliche und symbolische Repräsentationen über das Thema Fotosynthese zusammen. Kennzeichnend dafür sind Fotografien von Pflanzen, mikroskopische Aufnahmen von Spaltöffnungen und Blattquerschnitten sowie logische Bilder in Form von Liniendiagrammen zur Abhängigkeit der Fotosyntheserate von Außenfaktoren. Zusätzlich werden in diesem Teilgebiet der Biologie auch Fachtexte verwendet, die sich durch eine Vielzahl an Fachbegriffen wie z.B. Chloroplast, Lichtreaktion oder Calvin-Zyklus auszeichnen. Daneben spielen symbolische Repräsentationen für die Fotosynthese eine entscheidende Rolle, wenn es um Strukturformeln für Chlorophyll oder die Bruttogleichung der Fotosynthesereaktion geht. In einer wissenschaftlichen Hausarbeit von Jäger (2015) wurden Lernende aufgefordert, zu erläutern welche Repräsentationen im Biologieunterricht häufig vorkommen und was diese im Speziellen auszeichnen. Dabei nennen die Probanden, dass Bilder im Biologieunterricht häufig in Bezug zu anderen Bildern stehen und mehreres auf einmal bzw. immer vieles gleichzeitig zeigen. Des Weiteren wird angemerkt, dass viele Bilder als besonders typisch für die Biologie angesehen werden, z.B. ein Karyogramm. Im Hinblick auf MER beschreiben die Probanden vor allem verknüpfende Aspekte, dass beispielsweise eine bildliche Darstellung häufig nicht ohne die textliche Information genutzt werden kann und MER meist so vorliegen, dass Inhalte aufeinander bezogen sind. Zusätzlich erwähnen sie, dass im Biologieunterricht häufig prozessartige MER vorkommen, in denen eine typische Schritthaftigkeit, gefolgt von einer charakteristischen Verwendung von Pfeilen, dargestellt ist. Daraus lässt sich schlussfolgern, dass Darstellungsformen im Fach Biologie je nach vorliegendem Teilgebiet unterschiedlich sein können. Dies betrifft ihre Häufigkeit aber auch ihre zugrundeliegende Repräsentation selbst (Florian, Schmiemann, & Sandmann, 2015).

Vor diesem Hintergrund wird auch der Aufgabenkontext für die vorliegende Arbeit definiert. Entsprechend wird sich an inhaltlichen Teilgebieten der Biologie orientiert, die besonders charakteristische Vertreter biologiespezifischer Abbildungen enthalten (s. Kapitel 3.3.2). Somit wird der Aufgabenkontext auf die Anwendung eines Fachinhaltes innerhalb einer Testaufgabe beschränkt (vgl. Finkelstein, 2005). Demnach werden Kontexte als Frage- und Problemstellungen mit dem Merkmal „Kontextbezug“ verstanden (Finkelstein, 2005; Muckenfuß, 1995). Die so kontextualisierten Aufgaben sind damit inhaltlich im biologiespezifischen Fachwissen unterschiedlicher Teilgebiete wie z.B. der Ökologie verortet (s. Kapitel 3.3.2). Besonders relevant für das Abitur sind neben der Ökologie die Teilgebiete Physiologie, Zellbiologie, Genetik und Evolution (vgl. Florian, 2012; KMK, 2004). Solche typisch biologischen Aufgabenkontexte werden in authentische Alltagssituationen eingebettet und stellen den Fachinhalt innerhalb einer übergeordneten Aufgabenstellung dar (vgl. Finkelstein, 2005; van Vorst et al., 2014). Verbunden mit diesen fachspezifisch ausgewählten Kompetenzen zum Umgang mit MER, können in diesem Zusammenhang Auskünfte über die zur Lösung einer Aufgabe erforderlichen domänenspezifischen Anforderungen erfasst werden (vgl. Schecker & Parchmann, 2006).

1.6 ZUSAMMENFASSUNG

Die Biologie ist durch eine enorme Vielfalt an unterschiedlichen Darstellungsformen charakterisiert. Lernende müssen in der Lage sein, verschiedene Arten von Repräsentationen aufeinander beziehen zu können (Schnotz, 2002). Entscheidend ist nicht nur die Repräsentationsform, sondern auch inwieweit die Struktur der Darstellungsform zur repräsentierten textlichen Information ist (Saß, 2012; Schnotz & Bannert, 2003, s. Kapitel 1.4.1 und Kapitel 1.4.2). Dafür müssen Strukturmerkmale identifiziert und die Repräsentationstypen sinnvoll miteinander verknüpft werden, damit der Aufbau eines kohärenten mentalen Modells gelingt. Der Wissenserwerb und das Verständnis beim Lernen mit (multiplen) externen Repräsentationen ist somit entscheidend davon abhängig, inwieweit bestehendes Wissen mit den Informationen aus den Repräsentationen vernetzt und integriert wird (Schönborn & Bögeholz, 2009, s. Kapitel 1.4.3).

Die berichteten Studien fassen zusammen, dass im Unterricht der Umgang mit MER geübt werden muss (z.B. Bergey et al., 2015; Griffard, 2013; Roth & Pozzer-Ardenghi, 2013; Treagust & Tsui, 2013). Dies ist für Lernende bedeutsam, da MER und deren Integrationsleistung Gegenstand von Unterricht sind und dabei immer wieder Schwierigkeiten auftreten. Gerade das Herstellen von Zusammenhängen fällt Schülerinnen und Schülern schwer (Seufert, 2003). Haben Lernende ein geringes fachliches Vorwissen, zeigen sich häufig Probleme beim Identifizieren relevanter Informationen der dargestellten MER (z.B. Kozma, 2003). Aus diesem Grund ist es entscheidend, dass Schülerinnen und Schüler beim Umgang mit MER zunächst jede externe Repräsentation für sich verstehen und der Einsatz von MER nicht zum Selbstläufer wird. Trotz dieser kognitiv herausfordernden MER-Integration, belegen Studien, dass MER zu einem tieferen Verständnis von fachlichen Inhalten beitragen können (z.B. Hubber, Tytler & Haslam, 2010; Ainsworth, 2006). Gleichzeitig werden sie den nationalen Bildungsstandards gerecht, indem sie naturwissenschaftliche Inhalte als zusammenhängende Prozesse vermitteln und damit ein ganzheitliches konzeptuelles Verständniss naturwissenschaftlicher Themen fördern.

Vor diesem Hintergrund sind MER Gegenstand der vorliegenden Arbeit, indem sie eine Zeige- und Konstruktionsfunktion erfüllen. MER lenken die Aufmerksamkeit auf die kritischen Merkmale eines Gegenstands, während die Konstruktionsfunktion die bei der Bildung des mentalen Modells zentralen konzeptuellen Organisationen unterstützt (Weidenmann, 2002). Durch die Kombination von Text und Bild werden die jeweiligen Vorzüge miteinander verbunden. Texte besitzen eine höhere Ausdrucksmächtigkeit (Schnotz, 1994), in Bildern können Informationen oftmals ohne Umstände abgelesen und Inferenzen besser nachvollzogen werden

(Schnotz, 2003). Dadurch, dass MER unterschiedliche Codierungen verwenden und die zu erlernenden Inhalte auf verschiedene Modalitäten verteilen, kann eine mögliche Überlastung des Arbeitsgedächtnisses reduziert werden (s. Kapitel 2.2 und Kapitel 2.3). Für weitere Untersuchungen im Zusammenhang mit der Cognitive Load Theory (CTL) sei an dieser Stelle bereits auf die Arbeiten von Sweller, van Merriënboer, & Pass (1998), Sweller (1994), Sweller (2005) sowie Chandler & Sweller (1991) verwiesen.

2 THEORIEN UND MODELLE ZUM TEXT-BILD-VERSTEHEN

Lernen wird nicht als reines Wahrnehmen von Symbolen und ikonischen Zeichen verstanden, sondern als aktives Selektieren, Organisieren und Integrieren von Information, die von auditiven oder visuellen Reizen ausgehen (Kozma & Russell, 2005). Kommunikation mit Repräsentationen wird auch mit den Worten „convert our mental representations into external presentations“ (Rapp & Kurby, 2008, S. 33) beschrieben. Damit sind Repräsentationen ein kognitives Werkzeug und für das Lernen und Problemlösen essentiell (Schnotz & Bannert, 2003). Gerade multimediale Informationssysteme, d.h. die Kombination von Texten und Bildern (Weidenmann, 2002), steuern dabei unsere Verständnisprozesse. Bei der Verarbeitung solcher Text-Bild-Kombinationen werden schließlich interne Verarbeitungsprozesse ange-regt (Levin et al., 1987; Ainsworth, 1999). Damit ist die kognitive Verarbeitung ein zentraler Punkt beim Lernen mit multiplen externen Repräsentationen und Teil dieser Arbeit.

Das folgende Kapitel konzentriert sich auf das Text-Bild-Verstehen in Zusammenhang mit der Entwicklung von *Representational Competence*. Kognitionspsychologische Theorien und Modelle zum Text- und Bildverstehen werden an dieser Stelle nicht eingehend erläutert, sondern auf einschlägige Literatur verwiesen (Textverstehen vgl. van Dijk & Kintsch (1983); Kintsch (1998); Bildverstehen vgl. Weidenmann (1994); Pinker (1990)).

2.1 REPRÄSENTATIONSKOMPETENZ

Im deutschsprachigen Raum haben Wafi & Wirtz (2016) ein expertenbasiertes Strukturmodell von Visualisierungskompetenz vorgeschlagen. Darin wird zwischen einer rezeptiven, d.h. dem Erkennen, Verstehen einer Darstellung sowie dem Verknüpfen multipler Darstellungen und der produktiven Komponente, d.h. dem Generieren von Darstellungen, unterschieden. In diesem Modell werden fächerübergreifend Kompetenzen formuliert, die zu einer lernförderlichen Verarbeitung von Visualisierungen führen. Die einzelnen Kompetenzfacetten spiegeln dabei die bereits erwähnten Prozesse der Informationsentnahme, der Konstruktion und der Translation wieder. Die Übersetzungsleistung beim Umgang mit multiplen externen Repräsentationen wird unter dem internationalen Stichwort „*representational competence*“ zusammengefasst und wird definiert als „a set of skills and practices that allow a person to reflectively use a variety of representations or visualizations, singly and together, to think about, communicate, and act on chemical phenomena in terms of underlying, a perceptual physical entities and processes“ (Kozma & Russell, 2005, S. 131; Kozma & Russell, 1997).

Der Fokus richtet sich dabei auf die Verwendung von mehreren Repräsentationen zur Kommunikation. Dabei sind Fähigkeiten wie das Interpretieren, Konstruieren, Translatieren und Evaluieren ausgewählter Repräsentationen entscheidend (Nitz et al., 2014). Die Entwicklung von *Representational Competence* ist sowohl vom Kontext, von der Repräsentationsform selbst als auch vom Gebrauch in der vorhandenen Lehr-Lernsituation abhängig und nur über einen aktiven Umgang mit Repräsentationen zu erreichen (Hubber et al., 2010; Prain & Waldrip, 2006; Stieff, 2011; Waldrip, Prain, & Carolan, 2006; Kozma & Russell, 2005; Nitz, 2012). Entsprechend werden folgende Fähigkeiten zusammengefasst:

- I. Auswahl und Beurteilung von Repräsentationen hinsichtlich Angemessenheit und Zweckmäßigkeit (Kozma et al., 2000; Kozma & Russell, 1997).
- II. Koordination und Übersetzung unterschiedlicher Repräsentationsformen sowie ihre Beziehung zwischen ihnen erklären (Kozma et al., 2000; Ainsworth, 1999; Kozma & Russell, 1997).
- III. Merkmale, Eigenschaften sowie Informationen aus unterschiedlichen Repräsentationen identifizieren, beschreiben und mit anderen Repräsentationsformen vergleichen sowie neue Repräsentationen konstruieren (Kozma et al., 2000; Kozma & Russell, 1997).
- IV. Anwendung von Repräsentationen im Rahmen naturwissenschaftlicher Grundbildung, d.h. das Erkennen naturwissenschaftlicher Fragestellungen, Erklärung von Phänomenen sowie das Lösen von Problemen (Rönnebeck et al., 2006; Kozma & Russell, 2005; 1997; Kozma et al., 2000).

Kozma & Russell (2005) ergänzen für die Entwicklung von *Representational Competence* fünf weitere Komponenten. In einem ersten Schritt erfolgt das Erkennen der Oberflächenmerkmale. Anschließend werden über symbolische Fähigkeiten, syntaktische und folglich semantisch bedeutsame Repräsentationen gebildet. Als höchste Stufe nennen die Autoren den rhetorischen Gebrauch von Repräsentationen, wodurch die Fähigkeiten eine reflexive Komponente erhalten. Entscheidend sind auch hier wieder das Interpretieren, Konstruieren und Translatieren von multiplen externer Repräsentationen. Die Interpretation beinhaltet sowohl die Informationsentnahme aus einer oder jeder einzelnen Repräsentation als auch die Umwandlung von Information aus einer Repräsentation in eine andere (Mayer, 2002). Beim Konstruktionsprozess werden hingegen neue Repräsentationen selbstständig vom Lerner konstruiert, während Translationsprozesse ein Verständnis der Informationsverarbeitung und die Umwandlung zwischen verschiedenen Repräsentationen erfordern (Schönborn & Bögeholz, 2009; Ainsworth, 1999). Kozma & Russell (1997) beziehen in ihren Ausarbeitungen zusätzlich

die Organisationsebenen der Biologie in ihr Modell ein, d.h. sie unterscheiden zwischen subzellulärer, zellulärer und organischer Ebene. Schönborn & Anderson (2009) ergänzen die Repräsentationskompetenz durch die unterschiedlichen „modes of representation“, d.h. den Abstraktionsgrad einer Repräsentation, der bei Lernenden abhängig von der Vertrautheit mit den zugrundeliegenden Konventionen und der Fähigkeit des Vernetzens ist (Roth & Bowen, 2001; Ainsworth, 2006; 1999). Auf Basis neuerer Literatur und den von Ainsworth (1999) formulierten Funktionen von MER, entwickelten Tsui & Treagust (2013) ein Kompetenzmodell zum Umgang mit MER für den Biologieunterricht. Ihr ‚cube model‘ besteht aus drei Dimensionen (vgl. ABBILDUNG 3): *modes of representations*, *levels of representations*, und *domain knowledge of biology*.

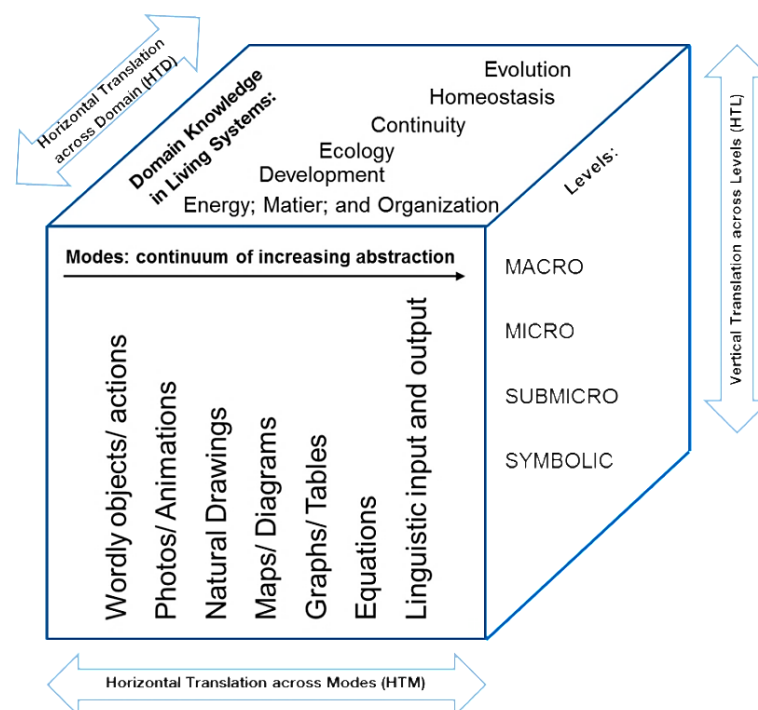


ABBILDUNG 3: 'THE CUBE MODEL' FOR LEARNING BIOLOGY WITH MER, VERÄNDERT NACH TSUI & TREAGUST, 2013, S. 10, FIG. 1.2

Unter *modes of representations* werden alle typisch naturwissenschaftlichen Darstellungsmodi zusammengefasst wie beispielsweise Zeichnungen, Fotografien, Animationen, Diagramme, Schemazeichnungen, die im Modell nach aufsteigender Abstraktion angeordnet sind (vgl. Pozzer & Roth, 2003 sowie Kapitel 1.4.2). Die *levels of representations* gehen den Autoren zufolge auf die von Johnstone (1982) formulierten Ebenen innerhalb der Chemie zurück und werden ebenfalls für die Biologie zugrunde gelegt. Für ein vollständiges Verstehen biologischer Phänomene sind hierzu vier verschiedene Organisationsebenen zu unterscheiden: die makroskopische, die zelluläre oder subzelluläre Ebene, die molekulare (submikroskopische) Ebene und die symbolische Ebene (Tsui & Treagust, 2013; Gilbert & Treagust,

2009). Den dritten Bereich innerhalb des Modells, das *domain knowledge of biology*, wurde in Anlehnung an die Biological Science Curriculum Study (2006) in verschiedene biologiespezifische, übergeordnete Bereiche untergliedert: Evolution, Homöostase, Energie, Kontinuität, Entwicklung und Ökologie. Mit diesem Modell formulierten die Autoren eine Basis, auf welcher Lernprozesse beim Umgang mit MER beschrieben werden können (Tsui & Treagust, 2013). Dabei geht es den Autoren vor allem um die Diagnose von Lernprozessen, die bei der Translation von MER stattfinden, da dies ein entscheidender Prozess bei der Entwicklung von *Representational Competence* ist. Hinzu kommt, dass in der Biologie ein vielfältiges Angebot unterschiedlicher Kategorien deklarativen Wissens vorliegt (vgl. Cheng & Gilbert, 2015; Bergey et al., 2015). In ihrer Studie berichten Nitz et al. (2014), dass sich im Verlauf einer Unterrichtsstunde das Fachwissen kontinuierlich verbessert, die Entwicklung von *Representational Competence* hingegen nur langsam ansteigt. Folgerichtig wird viel Zeit und Praxis benötigt, um Repräsentationskompetenz zu erwerben und sie expertenhaft zu nutzen, während das Präsentieren von Faktenwissen zu einem zunächst schnellen Wissenszuwachs führt (Kozma & Russell, 2005; Nitz et al., 2014). Die enorme Vielfalt biologiespezifischer Repräsentationen und ihre unterschiedlichen Abstraktionsgrade sind kognitiv herausfordernd (Ainsworth & Loizou, 2003; Ainsworth, 1999; Schönborn & Bögeholz, 2013). Diese hohe Komplexität von Biologieunterricht verlangt eine umfassende empirische Untersuchung, damit angehende und bereits unterrichtende Lehrkräfte eine entsprechende *Representational Competence* entwickeln und entsprechende, adressatengerechte Lehr-Lernkonzepte zum Umgang mit MER anbieten und vermitteln können (Eilam, 2012).

2.2 KOGNITIONSPSYCHOLOGISCHER HINTERGRUND

„Ein Bild sagt mehr als 1000 Worte“ (Barnard, 1921). Dieses Sprichwort lässt fälschlicherweise den Schluss zu, dass bildliche Darstellungen per se leicht zu verstehen seien. Dass ein Wissenserwerb allein durch Bilder und ohne weitere instruktionale Maßnahmen lernförderlich ist, widerlegen Studien (z.B. Ainsworth, 1999; 2006). Die Anwendung und der Umgang mit bildlichen Repräsentationen müssen explizit angesprochen werden (Schnotz, 1994). Text-Bild-Kombinationen und ihre Integration sind Gegenstand von Unterricht und Aufgabenkultur. Dabei müssen Schülerinnen und Schüler vielfältige Repräsentationen bei der Aufgabenlösung aufeinander beziehen und die Bedeutung der enthaltenden naturwissenschaftlichen Konzepte und Prinzipien erkennen. Sie müssen außerdem in der Lage sein, verschiedene Arten der Informationspräsentation, z.B. Fachwissen in Form eines Textes kombiniert mit einer abstrakten Repräsentation, aufeinander beziehen zu können (Schnotz, 2002). Somit stehen Lernende vor der Herausforderung Informationen aus den Informationsquellen Text und

Bild zu entnehmen und diese gleichzeitig sinnvoll zu verknüpfen, d.h. zu integrieren (McElvany et al., 2010). Dieses integrative Lesen ist in Schule und Unterricht grundlegend für Lernen und damit Gegenstand von Bildung und Forschung. Wie die kognitiven Prozesse dazu ablaufen, d.h. wie Informationen, die aus Texten und Bildern entnommen und zueinander in Beziehung gesetzt werden, kann in verschiedenen Modellen der kognitiven Verarbeitung von Repräsentationen gezeigt werden. Aus diesem Grund ist es notwendig, neben den vorgestellten Formen der externen Repräsentationen auch den Prozess der Informationsverarbeitung anhand von mentalen Repräsentationen zu erläutern.

Grundlage der kognitiven Verarbeitung externer Repräsentationen ist die Gedächtnisleistung bei der Informationsdarbietung. Zur Beschreibung der Gedächtnisleistung wird vor allem das Mehrspeichermodell von Atkinson & Shiffrin (1968) herangezogen, das in ABBILDUNG 4 dargestellt ist. Das Modell wurde im Laufe der Zeit ausdifferenziert und ergänzt, die Grundstruktur verlor jedoch nie an Gültigkeit und dient auch heute noch als Grundlage zur Informationsverarbeitung und ist Basis für Theorien und Modelle zur Verarbeitung von Repräsentationen.

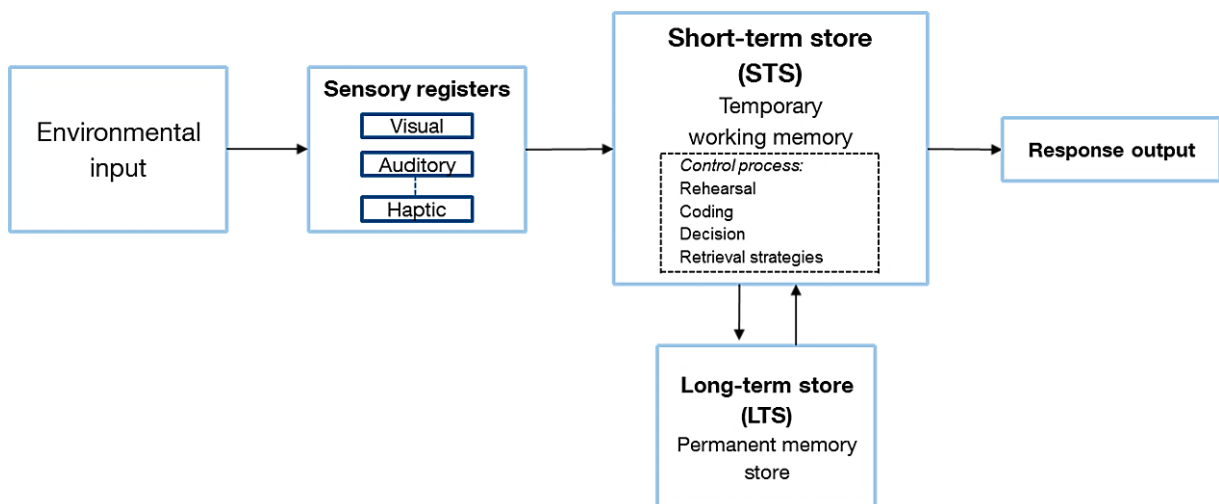


ABBILDUNG 4: MEHRSPICHERMODELL NACH ATKINSON & SHIFFRIN (1968), VERÄNDERT NACH ATKINSON & SHIFFRIN, 1971, S. 3B, FIGURE 1

Die Autoren ergänzten die bereits bestehenden funktionalen Speicher, das Kurzeit- (STS) und das Langzeitgedächtnis (LTS), um einen dritten Speicher, das sensorische Register (Ultra-kurzzeitgedächtnis). Dieses besteht aus verschiedenen Gedächtnissystemen, die wiederum modalitätsspezifisch sind. Alle Informationen, die aus der Umgebung aufgenommen und denen eine subjektive Bedeutung zugeschrieben werden, gelangen in das sensorische Register. Dabei werden vom sensorischen Register nur wichtige Informationen gefiltert, d.h. es dient als Puffer für die Fülle an Umgebungsinformationen. Relevante Informationen gelangen daraufhin in das Kurzzeitregister (Short-term store, STS), das einer Kapazitätsgrenze unterliegt

und nur wenige Inhalte gleichzeitig verarbeitet werden können (vgl. Millersche Zahl). Das Kurzzeitregister wird von Baddeley (1986; 2003) als Arbeitsgedächtnis bezeichnet und bezieht sich vor allem auf die aktive Verarbeitung von eingehenden auditiven und visuell-räumlichen Informationen aber auch von Informationen aus dem Langzeitgedächtnis. Baddeley (2003) nennt hierzu drei Systeme: Die phonologische Schleife zur Speicherung und Transformation von auditiver sowie sprachlicher Informationen, der räumlich-visuelle Notizblock zur temporären Speicherung von visuellen Wahrnehmungen und als drittes System die zentrale Exekutive zur Koordination und Kontrolle der erstgenannten Systeme. Das Modell wird von Baddeley (2003) durch den episodischen Puffer ergänzt, der sowohl visuelle als auch phonologische Informationen in Form von Episoden speichert. Dieses Modell stellt damit eine entscheidende Grundlage für das Verstehen der Verarbeitungsprozesse von MER im Arbeitsgedächtnis dar und ist für die Perzeption und Integration von MER im Rahmen einer Modellinspektion und Modellkonstruktion relevant (s. Kapitel 2.4). Das letzte Register ist das Langzeitgedächtnis (Long-term store, LTS), das zur permanenten Wissensspeicherung für Informationen dient. Die Einspeicherung von Wissen bedarf dabei wesentlich mehr Anstrengungen, da häufig nur elaboriertes Wissen eingespeichert werden kann. Im Unterschied zu den beiden anderen Registern ist hingegen die Kapazität und Speicherdauer nahezu unbegrenzt. Man geht davon aus, dass sogenannte kognitive Schemata die Informationseinheiten untereinander koordinieren und eine organisierte Wissensstruktur abbilden (Niegemann et al., 2008).

2.3 MULTIMEDIA-LEARNING

Die *Cognitive Theory of Multimedia Learning (CTML)* fokussiert sowohl die Verarbeitung von Text und Bildern als auch die Modalitätsunterschiede der Textpräsentation (Mayer, 2001; Schnotz, 2005). Dabei bedeutet der Begriff Multimedia, dass bei der Informationsvermittlung mehrere (meist apersonale) Mittler eingesetzt werden (Schnotz, 2001). Mayer (1997) definiert Multimedialernen, wenn Informationen in mehr als einer Codierung präsentiert werden und unterscheidet zwischen der Sinnesmodalität und dem Darstellungscode. Multimedia wird nach Weidenmann (2002) durch die drei Dimensionen Multimedialität, Multicodalität und Multimodalität definiert, wobei sich jeweils mono (,einzig') und multi (,vielfach') klassifizieren lassen. Die *CTML* geht auf Mayer (2001) zurück und zieht Annahmen aus Paivios *Dualer Codierungstheorie* (1986) als auch Ansätze aus der *Cognitive Load Theory (CTL)* sowie zum *Arbeitsgedächtnis* (Sweller & Chandler, 1994; Baddely, 1992) und der *Text-Bild-Integration* nach Schnotz (1994) heran. Paivio geht davon aus, dass unterschiedlich codierte Repräsentationen (verbale und visuell) in unterschiedlichen Kanälen des menschlichen Gedächtnisses

verarbeitet werden, d.h. die Verarbeitung von unterschiedlichen Darstellungsformen findet in zwei kognitiven Subsystemen statt. Diese Trennung ist auch Grundlage für Mayers Theorie (1997). Das Modell ist in ABBILDUNG 5 dargestellt und erlaubt Aussagen für die Gestaltung von multimedialen Lernapplikationen, die Mayer als „presentation involving word (such as spoken or printed text) and pictures (such as animation, video, illustrations, and photographs) in which the goal is to promote learning“ bezeichnet (Mayer, 2002, S. 55).

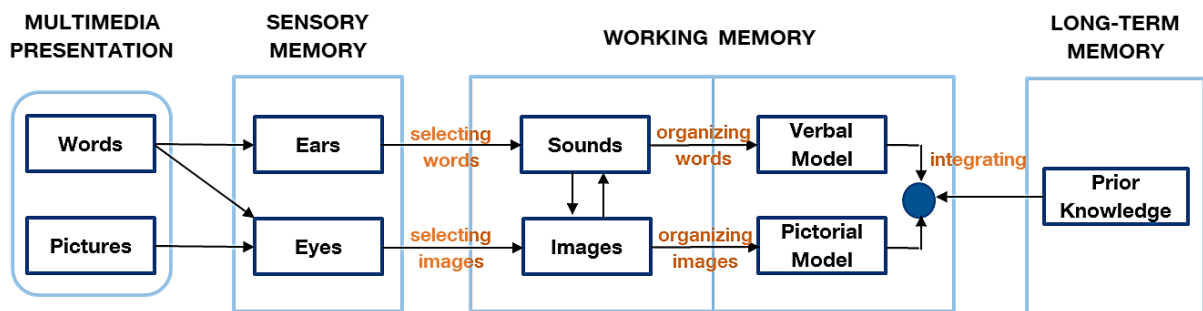


ABBILDUNG 5: KOGNITIVE THEORIE DES MULTIMEDIALERNEN NACH MAYER (1997), AUF BASIS VON MAYER, 2009, S. 61, FIGURE 3.1.

Der CTML von Mayer (2001; 2005; 2014) liegen drei Annahmen zugrunde. Die aktive Wissenskonstruktion wird von der Zwei-Kanal-Verarbeitung, von der ressourcenbezogenen Limitierung und drittens von aktiven Verarbeitungsprozessen bestimmt (Mayer, 1999; 2001; 2014; Mayer & Moreno, 2003; zitiert nach Unterbruner, 2007). Gemäß der Dual-Channel-Assumption werden zwei kognitive Subsysteme unterschieden. Das verbale System (für geschriebene oder gesprochene Texte) sowie das non-verbale System, d.h. für Bilder, Videos und Animationen, die als *representational modes* (Mayer & Moreno, 2003; Mayer, 2005) bezeichnet werden. Des Weiteren werden *sensory modalities* unterschieden, die den Sinnesmodalitäten (auditiv und visuell) zugewiesen werden (vgl. ABBILDUNG 5). Über multimediale Botschaften erreichen Wörter und Bilder das sensorische Gedächtnis, dies geschieht über Augen (geschriebene visuelle Texte sowie Bilder) und Ohren (gesprochene Texte). In einem sehr kurzen Zeitraum folgt der erste Verarbeitungsprozess, indem das sensorische Gedächtnis einen exakten visuellen und auditiven Eindruck speichert. Gemäß der *Limited-Capacity-Assumption*, die auf der *Cognitive Load Theory (CLT)* von Sweller (1994) basiert, wird auf die limitierte Menge der Informationsverarbeitung in den beiden kognitiven Systemen verwiesen (Mayer, 2005). In der *Active-Processing-Assumption* hält Mayer (2005) fest, dass Individuen für den Aufbau von kohärenten mentalen Repräsentationen aktive Verarbeitungsprozesse durchführen. Verstehendes und problemlösendes Lernen wird also immer dann ermöglicht, wenn ein Lerner kohärente mentale Modelle im auditiv-verbale und im visuell-piktoralen Kurzzeitgedächtnis bildet und diese Repräsentationen zu einem mentalen Modell integriert

werden, das mit dem Vorwissen abgeglichen wird (Mayer, 2002). Daraufhin folgt die bewusste und steuerbare Verarbeitung der Selektion, Organisation und Integration (Mayer, 2005). Im Anschluss an die Aufnahme und dem Transfer der Information in das Arbeitsgedächtnis folgt der erste Konstruktionsprozess, das *Selektieren*. Aufgrund der begrenzten Verarbeitungskapazität des Arbeitsspeichers ist es im ersten Konstruktionsprozess entscheidend, relevante Wörter und Bilder herauszufiltern (Mayer & Moreno, 2003; Mayer, 1997). Anschließend erfolgt im zweiten Konstruktionsprozess, dem *Organisieren*, die Umwandlung zu einem kohärenten mentalen Modell, das sich in ein verbales und ein piktorales Modell aufteilt (Mayer, 1997; 2002). Dabei wird angenommen, dass Lerner dabei unterschiedliche mentale Repräsentationen wie Oberflächenrepräsentationen, Propositionen des dargestellten Sachverhalts konstruieren (Schnotz, Bannert, & Seufert, 2002). Im dritten Schritt des Konstruktionsprozesses werden Verbindungen zwischen beiden codespezifischen Modellen gebildet, mit dem Vorwissen aus dem Langzeitgedächtnis abgeglichen und zu einem einheitlichen mentalen Modell *integriert* (Mayer & Moreno, 2003). Dieser Integrationsschritt erfordert ein effizientes Nutzen der kognitiven Ressourcen und verursacht eine hohe kognitive Belastung, ist aber entscheidend bei einer aktiven Konstruktion von Wissen und Grundlage für konstruktivistische Lernprozesse. Damit wird auch das Vorwissen zu einem entscheidenden Prädiktor. Es kann als „pre-instructional knowledge structures“ (Kluwe, 1988, S. 374, zitiert nach Gehl, 2013) verstanden werden und hat einen Einfluss auf das Rezeptionsergebnis (Ausubel, 1968). Da ausdifferenzierte Wissensstrukturen bereits vor einer Rezeptions-, Instruktions- oder Bearbeitungsphase vorhanden sein können, kann das benötigten Wissens zur Aufgabenlösung vorbestimmt sein und eine adäquate und objektive Beurteilung erschweren (Gehl, 2013). In den verschiedenen Studien wird darauf hingewiesen, dass das Vorwissen der Lernenden einer der stärksten Einflussfaktoren auf die Interpretationsleistung von Repräsentationen ist (vgl. Cook, 2006), da Vorwissen notwendige Verknüpfungen zwischen visuellen und verbalen mentalen Modellen erleichtert. Hegarty, Carpenter, & Just (1991) und Linn (2003) (vgl. Cook, 2006) konnten zeigen, dass Lernende mit geringem Vorwissen Schwierigkeiten haben zwischen relevanten und irrelevanten Informationen zu unterscheiden. Kozma (2003) sowie Kozma & Russell (1997) berichten, dass Lernende mit geringem Vorwissen häufig nicht dazu fähig sind, eine äquivalente Repräsentation für ein gegebenes Konzept zu finden. Vor diesem Hintergrund gilt es das Vorwissen als entscheidenden Einflussfaktor zu berücksichtigen und entsprechend zu kontrollieren (s. Kapitel 4.2).

Scheiter, Eitel, & Schüler (2016) berichten, dass sie die Annahme der *CTML* über kodaltätspezifische Repräsentationen nicht bestätigen konnten. In ihren Untersuchungen geht es um die Frage, zu welchem Zeitpunkt die Integrationsprozesse von Text- und Bildinformation

stattfinden und sich beeinflussen. Dabei konnten sie zeigen, dass Text- und Bildverarbeitungsprozesse bereits frühzeitig wechselseitig beeinflusst werden, wenn es um die Konstruktion eines integrierten mentalen Modells geht. Diese Befunde entsprechen den Annahmen von Schnotz (1999; 2001, s. Kapitel 2.4). Die Autoren weisen zusätzlich daraufhin, dass die Theorie der dualen Kodierung den Multimediaeffekt¹, der auf Erinnerungsleistungen beruht, nicht hinreichend erklären kann. Als Begründung führen sie an, dass der Textinhalt einen Einfluss auf den erinnerungsbezogenen Multimediaeffekt hat (Scheiter, Eitel, & Schüler, 2016; Schüler, Arndt, & Scheiter, 2015). Dies entspricht hingegen den Annahmen von Schnotz (1999; 2001), dass die Verarbeitung von Text und Bild ebenso voneinander abhängig sind wie die Konstruktionsprozesse bei kodalitätsspezifischer Gedächtnisrepräsentation im Arbeitsgedächtnis. Daneben konnten Scheiter & Eitel (2015) und Eitel & Scheiter (2015) zeigen, dass sich vor allem eine frühzeitige Bildbetrachtung positiv auf das Verstehen von Lerninhalten auswirkt und lernförderlich sein kann.

2.4 TEXT-BILD-INTEGRATIONSMODELL

Schnotz (1999; 2001; 2002) kritisiert am Modell zur *CTML*, dass von einer Verarbeitungsparallelität ausgegangen wird, indem sowohl für die Verarbeitung von Bildern als auch für die Verarbeitung von Texten eine duale Kodierung angenommen wird. Er verweist hingegen darauf, dass Texte und Bilder auf unterschiedlichen Symbolsystemen basieren und folglich verschiedene Repräsentationsprinzipien vorliegen. In diesem Sinne schlagen Schnotz & Bannert (1999) das integrative Modell zur Verarbeitung von Text und Bild (vgl. ABBILDUNG 6) vor und leisten damit einen entscheidenden Beitrag für das Verstehen von Integrationsprozessen beim Lernen mit multiplen externen Repräsentationen. Dabei beschreiben sie ebenfalls, welche Aspekte Lernvorteile und Probleme verursachen können. Damit multimediale Lernumgebungen verstanden und mit Vorwissen verknüpft werden können, ist eine zeitgleiche Verarbeitung aller Elemente der Text- und Bildinformation im Arbeitsgedächtnis notwendig, d.h. beim Bildverstehen werden multiple mentale Repräsentationen gebildet. Hierbei unterscheidet man Propositionen und mentale Modelle, wobei Propositionen als deskriptionale Repräsentationen und mentale Modelle als depiktionale Repräsentationen bezeichnet werden und eine Interaktion zwischen beiden stattfindet. (Schnotz, 2002; Schnotz, 2010).

¹ Unter dem Multimediaeffekt versteht man die Annahme, dass der Wissenserwerb bei der Kombination von Repräsentationen unterschiedlicher Kodalität (z.B. Text und Bild), die zusammen präsentiert werden, lernförderlicher ist als eine Repräsentation alleine (z.B. Mayer, 2001, 2009).

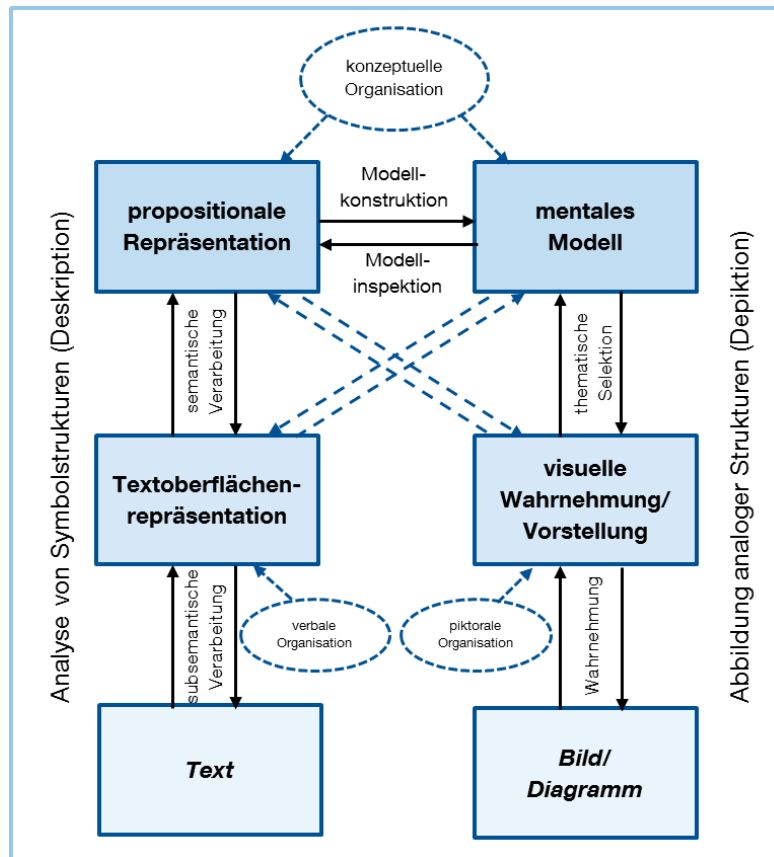


ABBILDUNG 6: INTEGRATIVES MODELL DES TEXT-BILD-VERSTEHENS NACH SCHNOTZ & BANNERT, 1999, AUF BASIS VON SCHNOTZ, 2001, S. 306, ABBILDUNG 1

Das integrative Modell des Text-Bild-Verstehens besteht aus einem deskriptionalen und einem depiktionalen Zweig und beschreibt wie diese unterschiedlichen Repräsentationstypen in verschiedenen Modalitäten (auditiv, visuell) verarbeitet werden. Im Modell können drei Ebenen unterschieden werden. Das visuelle sensorische Register, das visuelle Arbeitsgedächtnis sowie das Langzeitgedächtnis. Mit Pfeilen werden die Verarbeitungsprozesse (aufsteigend) und Selektionsprozesse (absteigend) symbolisiert. Querpfeile weisen auf Interaktionsprozesse im Arbeitsgedächtnis hin (vgl. ABBILDUNG 6). Die eingehende Information in Form von schriftlichem Text oder Bildern werden über das Auge in das visuelle sensorische Register aufgenommen und in das visuelle Arbeitsgedächtnis transportiert. Die Annahmen zur Verarbeitung von Informationen aus einem Text resultieren weitgehend aus der Textverstehensforschung (vgl. Kintsch & van Dijk, 1978). Dabei findet Textverstehen im verbalen Kanal statt, der relevante Informationen filtert und eine mentale Repräsentation der Textoberflächenstruktur in propositionale Repräsentationen überführt. Die Sprachinformationen werden gemäß syntaktischen und morphologischen Gesichtspunkten strukturiert, damit die gesamte sprachliche Information repräsentiert werden kann (Kintsch & van Dijk, 1978; Schnotz & Ban-

ner, 1999). Die Selektions- und Organisationsprozesse basieren auf Prozessen der Symbolverarbeitung und hängen von konzeptuellen bzw. inhaltlich kognitiven Schemata (Vorwissen) sowie Text-Schemata (Repräsentationswissen) ab (u.a. Schnotz, 2003). Anschließend findet über automatisierte Aufnahmeprozesse die perzeptive Encodierung statt, d.h. durch attentive Ablese- bzw. Inspektionsprozesse kann aus der propositionalen Repräsentation ein mentales Modell konstruiert oder elaboriert werden (Schnotz, 2001). Bilder werden im piktorialen Zweig verarbeitet. Dabei wird durch Wahrnehmungsprozesse und unter Einbeziehung kognitiver Schemata aus dem Langzeitgedächtnis eine interne mentale Repräsentation (visuelle Wahrnehmung/ Vorstellung) gebildet. Für das Verstehen der Depiktion folgen semantische Verarbeitungsprozesse, das aufgebaute mentale Modell dient dem Ablesen von Informationen, die wiederum in einer propositionalen Repräsentation gespeichert werden (Schnotz, 2005). Bei einer Text-Bild-Kombination werden analog der Verarbeitung einzelner Repräsentationen relevante Informationen der Sprachoberflächenrepräsentation in eine propositionale Repräsentation überführt. Auf der ersten Ebene findet das Ablesen von Detailinformationen statt. Daraus folgt die Verknüpfung von Text und Bild. Auf der zweiten Ebene werden einfache Relationen abgelesen, d.h. einfache semantische Relationen zwischen Text und Bild werden erkannt und auf die jeweils andere Repräsentationsform bezogen. Diese textbasierte propositionale Repräsentation wird zusammen mit der Bildoberflächenrepräsentation zur Konstruktion eines gemeinsamen mentalen Modells herangezogen (Schnotz, 2005). In dieser letzten Hierarchiestufe erfolgt die Integration der komplexen Relationsgefüge. Über Modellkonstruktions- und Modellinspektionsprozesse stehen das mentale Modell und die propositionale Repräsentation miteinander in Interaktion. Gleichzeitig können hierdurch Unterschiede in der Verarbeitung der drei Repräsentationstypen erklärt werden. So ist die propositionale Repräsentation Ausgangspunkt für die Erstellung eines mentalen Modells beim Textverstehen. Beim Bildverstehen werden ausgehend von mentalen Modell Informationen abgelesen und propositional encodiert (Schnotz, 1999). Beim Verstehen von Text-Bild-Kombinationen findet die Integration zu einer kohärenten Wissensstruktur über unterschiedliche, sich ergänzende Informationen von mentalem Modell und propositionaler Repräsentation statt (Schnotz, 2001). Dabei stellt das mentale Modell als Repräsentation die tiefste Form des Verständnisses dar, indem in diesem Text- und Bildrepräsentationen unterschiedlicher Modalitäten vereint werden (Kürschner & Schnotz, 2007). Abschließend ist anzumerken, dass alle Ebenen zueinander in Beziehung stehen. Dieser Aspekt ist vor allem dann entscheidend, wenn es um die Umwandlung von einer externen Repräsentation in eine andere Form der Darstellung geht oder die Integration von MER gefordert wird, wodurch multiple mentale Repräsentationen gebildet werden (Schnotz, 2002).

Vor diesem Hintergrund wurde auch das fachdidaktische Modell zur *Diagrammkompetenz* entwickelt (Lachmayer, 2008). Beim Einsatz von Diagrammen im Unterricht werden in Lernmaterialien häufig sowohl Text und ein Diagramm bereitgestellt. Lernende müssen dabei die Informationen aus dem Diagramm in Beziehung zu den Informationen aus dem Text setzen und die Informationen integrieren (Lachmayer, Nerdel, & Prechtel, 2007). Die Verarbeitung solcher integrativer Anforderungen kann ebenfalls mit dem integrativen Modell des Verstehens von Texten und Bildern erklärt werden (Schnotz & Bannert, 1999; Schnotz, 2001; Schnotz & Bannert, 2003). Werden anhand eines mentalen Modells Propositionen generiert, spricht man von Modellinspektion. Über die Modellkonstruktion können propositionale Informationen in das mentale Modell des Diagramms integriert werden. Die Integration umfasst damit sowohl Modellinspektions- und Modellkonstruktionsprozesse (Lachmayer, Nerdel, & Prechtel, 2007). Integrieren heißt, dass ein Lernender selbst erkennen muss, welche Informationen aus dem Diagramm abzulesen bzw. einzutragen sind und welche Ablese- und Konstruktionsprozesse kombiniert werden müssen. Kotzebue & Nerdel (2015) differenzieren zwischen der konstruktionsnahen Integration und informationsentnahen Integration von Diagrammen. Bei der konstruktionsnahen Integration geht es um das Identifizieren der relevanten Informationen in Text und Diagramm, um daraus Diagrammbestandteile zu konstruieren. Der Output ist entsprechend ein Bild. Die informationsentnahen Integration geht ebenfalls davon aus, zuerst relevante Informationen in Text und Diagramm zu identifizieren, fordert also auch eine aktive Verknüpfung von Text und Bild mit dem Resultat, dass ein Text geschrieben wird. Die Autoren konnten nachweisen, dass es sich bei beiden Konstrukten zwar um ähnliche, aber dennoch unterschiedliche Fähigkeiten handelt (Kotzebue & Nerdel, 2015).

2.5 ZUSAMMENFASSUNG

Die beschriebenen Theorien und Modelle zum Text- und Bildverstehen sowie von multimedialen Lernumgebungen liefern die notwendigen kognitionspsychologischen Grundlagen, wenn es um die Text-Bild-Integration und das Lernen mit MER geht. Sowohl die *Duale Kodierungstheorie* nach Paivio (1986), die *Cognitive Theory of Multimedia Learning (CTML)* nach Mayer (2005; 2014) als auch das *Text-Bild-Integrationsmodell* von Schnotz & Bannert (1999) zeigen, dass die Verknüpfung der unterschiedlich codierten Repräsentation entscheidend für Verstehensprozesse ist (s. Kapitel 2.2, Kapitel 2.3 und Kapitel 2.4).

Die Wissensvermittlung erfolgt dabei über unterschiedliche Zeichensysteme wie piktorale, verbale oder zahlengebundene Darstellungen, wobei Informationen verschiedenartig enkodiert werden (Weidenmann, 2002; Schnotz, 2001). Aufgrund der Vielzahl unterschiedlicher Repräsentationen und deren Kombination als MER, die im Unterricht Gegenstand von Arbeits- und Lernmaterialien sind, ist eine Kenntnis über die zugrundeliegenden Verarbeitungsprozesse entscheidend. So können über MER Sachverhalte auf allen Organisationsebenen der Biologie sowie unterschiedlichen Perspektiven dargestellt. Eine kohärente Wissensstruktur kann jedoch nur durch die Verknüpfung und das wechselseitige Übersetzen gebildet werden. Entscheidend dabei ist, dass die kognitionspsychologischen Grundlagen berücksichtigt und entsprechende Hilfestellungen für die Übersetzungsleistung angeboten werden, wenn der Repräsentationswechsel gelingen soll (Ainsworth, 2006). Vor diesem Hintergrund treten unterschiedliche repräsentationsspezifische Kompetenzen in den Fokus der vorliegenden Arbeit. So werden neben der Integration durch Informationsentnahme und Interpretation, auch die konstruktionsnahe Integration sowie Integrationsprozesse beim Translatieren und Transformieren zu und aus mehreren Repräsentationen untersucht. Um diese Zusammenhänge zu untersuchen bedarf es weiterführender Repräsentationsforschung, die biologiespezifische Text-Bild-Kombinationen untersucht (s. Kapitel 3.3).

3 KOMPETENZMODELL ZUR INTEGRATION VON MER

Auf Basis der theoretischen Betrachtung von Text-Bild-Kombinationen und ihrer kognitiven Verarbeitung in Zusammenhang mit aktueller Forschung zum Lernen mit MER als Beitrag für Repräsentationskompetenz, tritt der Begriff der Kompetenz in den Fokus. Dies gilt insbesondere für den Erwerb von fachsprachlicher Kompetenz, wenn es um den Umgang mit unterschiedlichen Darstellungsformen geht. Aktuell lassen sich in der Literatur keine adäquaten Kompetenzmodelle zum Umgang mit MER und deren Integration im naturwissenschaftlichen Unterricht finden. Aus diesem Grund werden in diesem Kapitel Theorien und Vorgehensweisen der Kompetenzmodellierung vorgestellt und die theoretische Entwicklung des Kompetenzstrukturmodells zum Umgang mit MER dargelegt.

3.1 KOMPETENZBEGRIFF

Der Kompetenzbegriff wird in der Fachliteratur sehr heterogen definiert. Diese Vielfalt verlangt eine begründete Abgrenzung des Begriffs Kompetenz für die vorliegende Arbeit, um für die Entwicklung eines Kompetenzmodells eine theoretisch fundierte Basis zu sichern. Im folgenden Forschungskontext werden Kompetenzen als „kontextspezifische Leistungsdispositionen unter Ausschluss motivationaler und affektiver Faktoren“ (Hartig & Klieme, 2006, S. 129) verstanden. Diese Definition fokussiert damit den kognitiven Anteil einer Kompetenz nach der Definition von Weinert (2001) und berücksichtigt die Handlungsfähigkeit von Personen in unterschiedlichen Kontexten als Voraussetzung für die Partizipation an gesellschaftlichen Diskussionen und Grundlage von lebenslangem Lernen. Grund für diese Fokussierung ist die Forderung, dass motivationale von kognitiven Aspekten getrennt zu erfassen sind (Weinert, 2001), da sie erst so eine empirische Untersuchung zugänglich machen können. In dieser Definition wird zusätzlich die Kontextspezifität betont. Der Kompetenzbegriff wurde bereits während seiner historischen Genese auf konkrete Anforderungssituationen bezogen (Klieme & Leutner, 2006; Schecker & Parchmann, 2006; Connell et al., 2003; White, 1959). In ihrer Definition sprechen Klieme & Leutner (2006) darüber hinaus davon, dass sich kontextspezifische kognitive Leistungsdispositionen „funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen“ (S. 4), d.h. dass die Formulierung und Erfassung von Kompetenzen einen zusätzlichen fachspezifischen Anteil hat. Speziell für die Biologie sollte die Definition von Kompetenz damit noch differenzierter gewählt werden. Um einer lebensnahen Ausrichtung wie der Biologie gerecht zu werden, müssen neben konzeptuellen auch biologiebezugene Aspekte berücksichtigt werden (Kampa, 2012). Demnach können Kompetenzen

in der Biologie als „erlernbare kognitive, kontextspezifische Leistungsdispositionen [verstanden werden], die (...) dazu befähigen, biologiebezogene Probleme in variablen Situationen zu lösen“ (Kampa, 2012, S. 43). Dementsprechend gilt es neben dem deklarativen biologischen Fachwissen auch das prozedurale Wissen in Form von Handlungen zu berücksichtigen, indem z.B. kommunikative Kompetenzen ausdrücklich gefordert und im Biologieunterricht thematisiert werden (KMK, 2005a, s. a. Kapitel 1.5).

3.2 KOMPETENZMODELLIERUNG

Die Diagnostik von Kompetenzen spielt bei Entscheidungen über Auswahl und Förderung, in Evaluationsprozessen als auch in der Grundlagenforschung eine besondere Rolle. Kompetenzmodelle stehen dabei zwischen Forschung und Unterrichtspraxis (Bernholt, Parchmann, & Commons, 2009). Ziel ist es entwickelte Kompetenzstrukturen empirisch abzusichern, um evidenzbasierte (biologie-) didaktische Entscheidungen zur Gestaltung von Aufgaben, Lernmaterial und Unterricht mit Blick auf die Fähigkeiten von Schülerinnen und Schülern zu treffen. Dabei ist die Erfassung von Kompetenzen ein theoretisch und methodisch komplexer Prozess in der empirischen Bildungsforschung, wenn diese Modelle Ausgangspunkt für die Entwicklung adäquater Messverfahren sein sollen (Hartig & Klieme, 2006).

Unter Kompetenzmodellierung versteht Rost (2006) den Prozess, der zur Formulierung eines Kompetenzmodells führt. Kompetenzmodelle beschreiben dabei „theoretische Beschreibungen der Struktur spezifischer Fähigkeiten, [die] als Antwort auf typisch moderne Problemlagen konzipiert sind“ (Klieme et al., 2003, S. 65). Dabei kann bei einem explorativen Vorgehen zunächst nach Mustern in bereits bestehenden Daten gesucht werden. Werden bei einzelnen Messungen hohe Zusammenhänge untereinander gefunden, können diese in Dimensionen zusammengefasst werden, d.h. sie erfassen dasselbe Merkmal. Werden hingegen nur geringe Interkorrelationen gefunden, so ist davon auszugehen, dass diese Messungen unterschiedlichen Merkmalen entsprechen (Hartig & Klieme, 2006). Das Ergebnis eines solchen deskriptiven Vorgehens führt zu einem deskriptiven Kompetenzmodell. Als zweiten Weg zur Formulierung eines Kompetenzmodells führt Rost (2006) das normative Vorgehen auf. Für die Kompetenzmodellierung wird auf relevante kognitionspsychologische Modelle Bezug genommen und der aktuelle Forschungsstand berücksichtigt. Die Kompetenzen lassen sich in einer Art Matrixstruktur abbilden, die wiederum Stufen enthalten kann. Die empirische Überprüfung solcher Kompetenzmodelle erfolgt über Testaufgaben, die dem postulierten Modell zugeordnet werden können. Items werden entwickelt, in einer Vorstudie eingesetzt und aufgrund statistischer Kennwerte optimiert. In der Hauptstudie erfolgt schließlich die Überprü-

fung des Kompetenzmodells (Rost, 2006). Zusammengefasst dient die empirische Validierung eines entwickelten Kompetenzmodells der Überprüfung der eingesetzten Items, inwieweit diese das Kompetenzmodell adäquat repräsentieren, sowie als Grundlage für die Interpretation einer entsprechenden Kompetenz (Terzer, Hartig, & Upmeyer zu Belzen, 2013).

In diesem Zusammenhang können verschiedene theoretische Kompetenzmodelle herangezogen werden. Theoretische Modelle können einerseits Strukturen definieren oder Niveaus charakterisieren (Klieme & Leutner, 2006). Damit können Kompetenzen unterschiedlich kognitiv modelliert und voneinander abgegrenzt werden. Kompetenzmodelle dienen dazu, die Erwartungen bezüglich der Lernergebnisse von Lernenden einer bestimmten Altersstufe in einem spezifischen Fach zu beschreiben sowie nach empirischen Erkenntnissen aufzuzeigen, wie diese Kompetenz erlangt werden kann (Klieme et al., 2003). Kompetenzmodelle fokussieren also eine theoretische Fundierung sowie empirische Beschreibung naturwissenschaftlicher Kompetenzen (Mayer & Wellnitz, 2014) und sind Ausgangspunkt für die Entwicklung von psychometrischen Modellen und schließlich von Messverfahren (Klieme & Leutner, 2006). Ausgangspunkt für die Entwicklung und Modellierung eines Kompetenzmodells sind in der Regel die Inhalte eines Faches im Rahmen der Curricula oder angestrebte Bildungsziele in Form von Kompetenzen wie sie die Bildungsstandards definieren. Diese wurden jedoch auf normativer Basis entwickelt, d.h. es fehlt ihnen sowohl eine theoretische Basis als auch eine empirische Absicherung (Köller, 2008). Sie dienen jedoch als Grundlage für die Operationalisierung von Bildungszielen und erlauben dadurch eine Vermittlung zwischen Output-Steuerung und konkreten Aufgabenstellungen (Klieme et al., 2003). Es zeigt sich jedoch, dass die Verbindung von Theorie und Empirie stets eine inhaltlich-fachliche Herausforderung der Kompetenzmessung ist und bisher entwickelte Modelle immer noch einen hohen Grad an Abstraktion aufweisen (Klieme & Leutner, 2006). Im Folgenden sollen verschiedene klassifizierte Kompetenzmodelle kurz vorgestellt werden, die Einordnung erfolgt nach normativen und deskriptiven Modellen (vgl. Rost, 2006; Schecker & Parchmann, 2006).

3.2.1 KOMPETENZSTRUKTURMODELLE

Bei der Modellierung von Kompetenzstrukturmodellen geht es darum, verschiedene Dimensionen von Kompetenz in einem spezifischen Bereich zu differenzieren (Klieme & Leutner, 2006), d.h. ein nach Dimension gegliedertes Gefüge von Teilkompetenzen zu gestalten. Dabei gilt es zu entscheiden, welche Kompetenzen sich anhand welcher Inhalte erheben lassen (Hartig et al., 2008). Schecker & Parchmann (2006) unterscheiden normative und deskriptive Kompetenzstrukturmodelle. Normative Modelle beschreiben die „(kognitiven) Voraussetzungen, über die ein Lernender verfügen soll, um Aufgaben und Probleme in einem bestimmten

Gegenstands- oder Anforderungsbereich lösen (zu) können“ (Schecker & Parchmann, 2006, S. 47). Grundlage solcher Modelle sind Theorien aus der Lernpsychologie. Damit normative Modelle, wie sie bei den Bildungsstandards vorliegen, als Grundlage für die Operationalisierung eines Kompetenzstrukturmodells herangezogen und damit empirisch evaluiert werden können, müssen die angestrebten Lernergebnisse beschrieben und geeignete Testverfahren ausgewählt werden (Klieme et al., 2003). Die Entwicklung von Kompetenzstrukturmodellen kann in vier Schritten erfolgen (z.B. Klieme & Leutner, 2006; Mayer & Wellnitz, 2014). Im ersten Schritt geht es darum, die formulierten Kompetenzen oder Standards in Kompetenzkonstrukte zu überführen, die theoretisch und empirisch abgesichert sind. Präskriptiv gesehen, geht es bei der Anzahl der Dimensionen um die Strukturelemente bzw. die Komponenten, die im Rahmen des Lehrangebots und der erwarteten Lernergebnisse stehen (Schecker & Parchmann, 2006). Inhaltlich gilt es durch Zusammenhangsstrukturen zu bestimmen, welche Kompetenzen differenziert erfasst werden sollen und auch aus diagnostisch-pragmatischer Sicht gerechtfertigt sind (Hartig & Klieme, 2006). Dadurch kann schließlich eingeschätzt werden, welche Korrelationen zwischen Testleistung und den unterschiedlichen Anforderungen vorliegen und auf welchen Dimensionen Unterschiede innerhalb des Kompetenzkonstrukts beschrieben werden können. Im Anschluss daran folgen methodische Schritte der Kompetenzentwicklung, d.h. die Abbildung der zu erfassenden Konstrukte in ein adäquates Messmodell, die Entwicklung der Messinstrumente sowie die anschließende empirische Überprüfung.

Im Unterschied dazu beschreiben deskriptive Modelle ein so genanntes Muster von (kognitiven) Voraussetzungen, mit dem man versucht das Lösungsverhalten der Lernenden zu rekonstruieren. Hierfür ist eine empirische Evidenz aus Fachdidaktik und Lernpsychologie erforderlich. Viele dieser deskriptiven Kompetenzmodelle wurden im Rahmen von großen Schulleistungsstudien wie TIMSS und PISA entwickelt und fokussieren vor allem allgemeine, naturwissenschaftliche Kompetenzen, die fächerübergreifend zu verstehen sind. Damit sind solche Modelle allerdings nicht auf die einzelnen Fächer übertragbar, d.h. „für die einzelnen Fächer (müssen) Kompetenzmodell entwickelt werden“ (Klieme & Steinert, 2004, S. 133). Gleichzeitig ist die Spezifität für ausgewählte Teilpopulationen problematisch, indem häufig nur einzelne Klassenstufen, Schulformen oder eine bestimmte Alterskohorte untersucht wurde (Schmiemann, 2010). Die Generalisierung solcher Ergebnisse sollte hingegen vorsichtig interpretiert und angenommen werden (Rost, 2004).

3.2.2 KOMPETENZNIVEAUMODELLE

Sollen Kompetenzniveaus beschrieben werden, „geht es darum, welche konkreten situativen Anforderungen Personen bei welcher Ausprägung einer Kompetenz bewältigen können“ (Klieme & Leutner, 2006, S. 7). Damit sind qualitative Kompetenzabstufungen gemeint, die eine Kompetenz graduieren können (Mayer & Wellnitz, 2014). Dabei geht es um die Frage, welche spezifischen Anforderungen eine besonders fähige Person bewältigt und eine Person mit niedriger Kompetenz gerade noch bewältigen kann und welche nicht (Hartig & Klieme, 2006). Kompetenzniveaumodelle versuchen also herauszufinden, über welche spezifischen Kompetenzen Lernende verfügen und welche Leistungsanforderungen in welcher Güte sie diese bewältigen können. Hierfür ist eine kriteriumsorientierte Testwertinterpretation nötig (Hartig & Klieme, 2006; Klauer, 1986). Eine kontinuierliche Skala der quantitativen Leistungswerte wird in Abschnitte unterteilt. Diese Abschnitte können dann als Kompetenzniveaus bezeichnet werden. Klieme et al. (2003) fordern, dass Kompetenzmodelle eine hierarchisch gegliederte Dimension enthalten sollten, um Unterschiede in der Kompetenzausprägung zu beschreiben. Hartig (2007) weist bei der Beschreibung von Kompetenzniveaus auf zwei Merkmale hin. Zum einen geht es um den Differenzierungsgrad der Kompetenzniveaus, der wiederum von der theoretischen Fundierung und den Daten abhängig ist. Zum anderen ist der Zeitpunkt der Kompetenzniveaubeschreibung entscheidend. Dies kann einerseits a priori passieren, d.h. auf Basis einer Theorie und damit vor der Datenerhebung. Damit kann bereits vorher geschätzt werden, welche Kompetenzstufe notwendig ist, damit die Aufgabe erfolgreich gelöst werden kann (Klieme et al., 2003). Die Niveaubeschreibung kann aber auch post hoc erfolgen, d.h. Ausgangspunkt sind die empirisch gewonnenen Daten nach der Erhebung (Mayer & Wellnitz, 2014). Eine weitere Möglichkeit ist die theoriebasierte Operationalisierung der Graduierung vor der Erhebung und die anschließende Beschreibung möglicher Niveaus auf Basis der empirischen Daten (vgl. z.B. ESNaS, Wellnitz et al., 2012). Im Zusammenhang mit Kompetenzstrukturmodellen wird damit zunächst eine Kompetenz in Kompetenzdimensionen strukturiert, bevor jede der ermittelten Kompetenzdimensionen wieder in Kompetenzniveaus unterteilt wird. Diese Modellierung basiert dabei auf inter- und intraindividuellen Unterschieden zwischen den Personen.

3.2.3 KOMPETENZENTWICKLUNGSMODELLE

Kompetenzentwicklungsmodellen liegt eine zeitliche Perspektive zugrunde, sie beantworten die Frage in welcher Weise Kompetenzen nach Altersstufen erreicht werden und welche Kompetenzstrukturen sich dadurch herausbilden können, d.h. wie sich Kompetenzen im zeitlichen Verlauf verändern und entwickeln (Schecker & Parchmann, 2006; Neumann et al.,

2007). Gleichzeitig wird die innere Dynamik der Kompetenzentwicklung in einer strukturellen Perspektive untersucht. Die Herausforderung besteht dabei in der Verknüpfung von fachlicher Perspektive und situativen Einflussfaktoren (Schecker & Parchmann, 2006), d.h. es geht um den Aufbau von Kompetenzentwicklungsmodellen, die „den diagnostizierten Kompetenzstand einer Schülerin/ eines Schülers mit instruktionsrelevanten Informationen (...) verknüpfen“ (Bernholt, Parchmann, & Commons, 2009, S. 221). Kompetenzentwicklungsmodelle sind empirisch fundiert und können dadurch einen entscheidenden Beitrag für die kumulative Förderung von Kompetenzen leisten (Hammann, 2004). Sie greifen über Jahrgangsstufen hinweg bestehende Kompetenzniveaus auf und bieten den Ausgangspunkt dafür diese weiterzuentwickeln. Dabei bieten sie zusätzlich eine Rückmeldefunktion von Kompetenzzuwachs und sind eine wichtige Grundlage für Lehr-Lernprozesse im Unterricht (Neumann et al., 2007). Der bisherige Forschungsstand zeigt jedoch, dass empirische Studien zu Kompetenzentwicklungsmodellen bisher kaum vertreten sind und vor allem validierte Modelle bisher nur in geringer Zahl vorliegen (Ferber, 2014).

3.3 KOMPETENZSTRUKTURMODELL ZUR INTEGRATION VON MER

Fachspezifische Kommunikation verläuft unter Verwendung von Fachbegriffen und Darstellungsformen. Dies gilt insbesondere für den Erwerb wissenschaftlicher Kompetenzen, wenn fachspezifische Repräsentationen wie z.B. Chromatogramme oder das Foto einer Gelelektrophorese im Biologiestudium interpretiert werden müssen. Dabei sind MER häufig nicht intuitiv verständlich, erfordern Lese- und Übersetzungsfähigkeiten und bereiten Lernenden unterschiedlicher Altersstufen Schwierigkeiten (Schnotz, 2002; von Kotzebue, Gerstl, & Nerdel, 2015; s. Kapitel 1.4.3). Spezifikationen der Repräsentationskompetenz werden aktuell meist lediglich auf Basis von Einzelpräsentationen vorgenommen. Es fehlen insbesondere für die Biologie Kompetenzmodelle zur Integration von MER, die Aussagen über die Schwierigkeiten verschiedener Repräsentationswechsel liefern, die zugrundeliegenden Gestaltungsmerkmale und Konventionen sowie den biologischen Kontext berücksichtigen. Zu diesem Zweck wurden internationale und nationale Standards, Curricula sowie Schulleistungsstudien gesichtet und bereits bestehende, fachdidaktische und allgemeinspsychologische Theorien und Modelle analysiert (vgl. Mayer & Wellnitz, 2014). Dadurch kann ein nomologisches Netzwerk aufgespannt werden, damit „die Gesamtheit der auf theoretischer Ebene angenommen Zusammenhänge zwischen dem interessierenden Konstrukt und anderen theoretischen Konstrukten“ (Hartig & Jude, 2007, S. 22) abgebildet wird.

Für das Kompetenzstrukturmodell werden drei Dimensionen postuliert, die sich aufgrund der theoretischen Herleitung voneinander abgrenzen lassen. Im Fokus des Modells stehen die kognitiven Prozesse in Form der Integration von MER sowie biologiespezifische Inhalte und unterschiedlichen Repräsentationstypen. Dadurch werden sowohl kognitive als auch inhaltsbezogene Merkmale einbezogen, um den kontextualisierten Charakter von Kompetenzen zu berücksichtigen und einen Beitrag zur Konstruktvalidierung zu leisten (vgl. Hartig & Frey, 2012). Folgende Dimensionen werden zugrunde gelegt (vgl. ABBILDUNG 7) und anschließend ausführlich beschrieben (s. Kapitel 3.3.1, Kapitel 3.3.2 und Kapitel 3.3.3).

- I. Die Dimension *Integration multipler externer Repräsentationen (MER)* beschreibt das Wissen über zentrale Integrationsprozesse beim Lösen von Aufgaben mit Repräsentationen und umfasst drei Teilkomponenten, die sich je nach Integrationsprozess und Output unterscheiden. Die zu untersuchenden Integrationsprozesse sind *Informationsentnahmenähe Integration (MERI)*, *Konstruktionsnahe Integration (MERII)* sowie *Translation und Transformation (MERIII)*.
- II. Die Dimension *Aufgabenkontext* richtet sich an verschiedene Themengebiete innerhalb der Domäne Biologie. Sie wird unterteilt in die drei Teilkomponenten *Ökologie*, *Stoffwechselphysiologie* und *Genetik*.
- III. Die dritte Dimension *Repräsentationstyps* folgt aus den heterogen vorliegenden Befunden zu biologiespezifischen logischen Bildern und Schemata. In dieser Dimension wird zwischen den Teilkomponenten *Diagrammen* und *Schemazeichnungen* unterschiedlicher Abstraktionsgrade differenziert.

Die Kombination jeweils einer der genannten Teilkomponenten aller Dimensionen ergibt eine konkrete Kompetenzanforderung. Dabei wird auf eine repräsentationsspezifische Trennung der Teilkomponenten geachtet, um eine Konfundierung der Variablen auszuschließen und um die Kompetenzen systematisch erfassen zu können (vgl. Prenzel et al., 2002). Das bedeutet, dass sich die erste Dimension ausschließlich auf Integrationsprozesse bezieht, während die zweite Dimension verschiedene biologische Inhalte als Komponenten unterscheidet. Die dritte Dimension berücksichtigt unterschiedliche Repräsentationstypen (vgl. ABBILDUNG 7). Das Modell konzentriert sich auf biologiespezifische, kommunikative Kompetenzen in unterschiedlichen Aufgabenkontexten. Dabei fokussiert es den Umgang mit biologiespezifischen Repräsentationen als Aspekt fachlicher Kommunikationskompetenz wie sie in den Bildungsstandards definiert ist. Das entwickelte Kompetenzmodell kann in bestehende Rahmenkonzeptionen eingeordnet werden (KMK, 2005a, 2004) und zeigt eine nationale (Lachmayer, 2008; von Kotzebue, 2014) sowie internationale Anschlussfähigkeit mit Blick auf *Representational Competence* (u.a. Tsui & Treagust, 2013; Anderson et al., 2013).

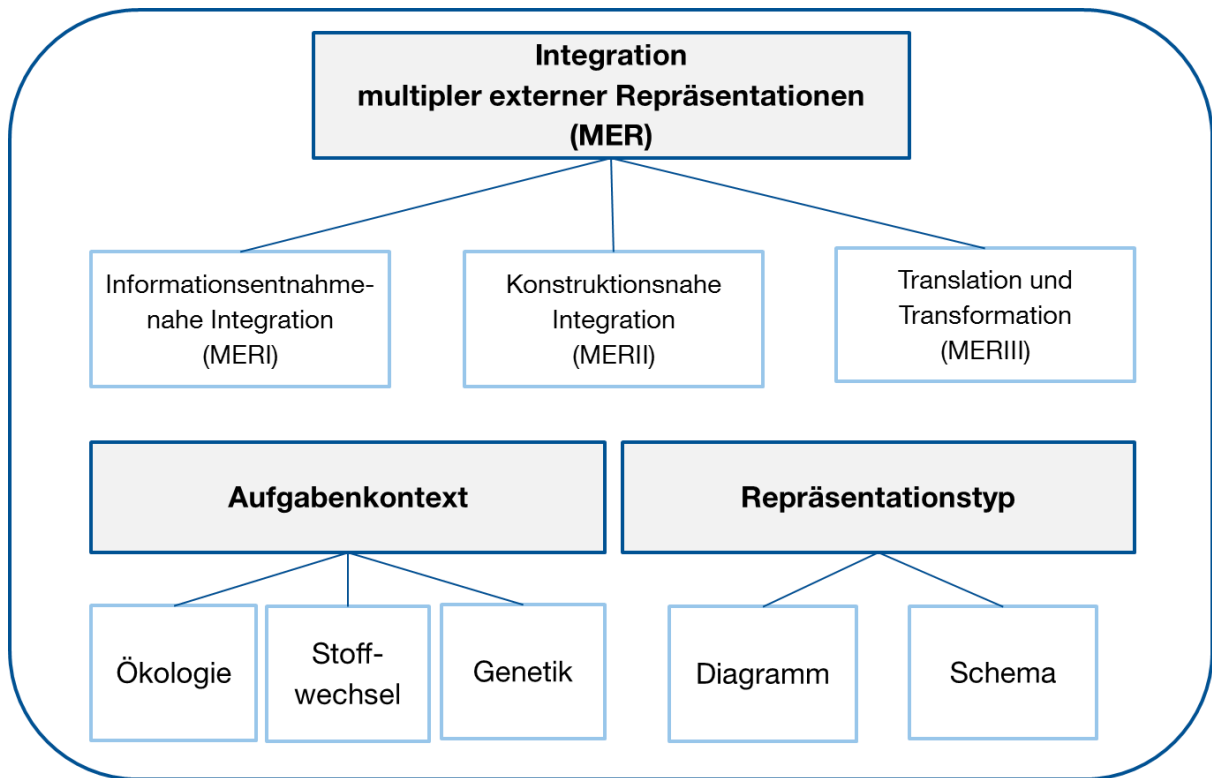


ABBILDUNG 7: KOMPETENZSTRUKTURMODELL ZUM UMGANG MIT MULTIPLLEN EXTERNEN REPRÄSENTATIONEN (MER) IM BIOLOGISCHEN KONTEXT

3.3.1 DIMENSION UMGANG MIT MULTIPLLEN EXTERNEN REPRÄSENTATIONEN

Die standardisierte Videoanalyse von Jatzwauk, Rumann, & Sandmann (2008) konnte zeigen, dass der Biologieunterricht vor allem Aufgaben mit kurzen Antworten enthält. Verbalisierungen oder bildhafte Darstellungen werden weniger verlangt. Problematisch in diesem Zusammenhang ist, dass Lern- und Leistungsaufgaben häufig Text-Bild-Kombinationen in Form von ER und MER enthalten. Gerade in Abituraufgaben wird in einer Analyse von Florian (2012) dargelegt, dass bei mehr als 42% der untersuchten Aufgaben aus den Jahren 2008 und 2009 als Lösungsweg eine Darstellung bzw. Konstruktion verlangt wurde. Die Anforderung besteht darin, die gegebenen Informationen zu erschließen und zu interpretieren sowie darin, dass Lernende solche Darstellungen selbst konstruieren müssen. Im Rahmen einer qualitativen Studie von Maier et al. (2010) wurde ein Kategoriensystem für die fächerübergreifende Aufgabenanalyse entwickelt und erprobt. Darin werden Repräsentationsformen des Wissens berücksichtigt und in *Aufgaben mit einer Repräsentationsform*, *Integration verschiedener Repräsentationsformen* und *Integration und Transformation des Wissens* unterschieden. Im biologiespezifischen Modell von Tsui & Treagust (2013) fokussieren die Autoren ebenfalls sowohl idealtypische Darstellungsformen als auch unterschiedliche Bereiche der Biologie. Zusätzlich gehen in das Modell die Organisationsebenen der Biologie ein (s. Kapitel 2.1).

Diese Modelle bieten einen Anknüpfungspunkt, weitere Kompetenzkonstrukte zum Umgang mit MER zu definieren, die fachdidaktisch relevant und den Integrationsprozess bei der Verwendung verschiedener Repräsentationstypen beleuchten. Auf der Basis der vorgestellten Theorien zum Text-Bild-Verständnis sowie von *Representational Competence* wird die Dimension *Integration multipler externer Repräsentationen (MER)* definiert und folgende Teilkomponenten in das Kompetenzstrukturmodell integriert: *Informationsentnahmenähe Integration (MERI)*, *Konstruktionsnahe Integration (MERII)* sowie *Translation und Transformation aus mehreren Repräsentationen (MERIII)*.

Die ersten beiden Komponenten gehen auf die Untersuchungen zur Diagrammkompetenz von Lachmayer (2008) und von Kotzebue (2014) zurück. Der Prozess der Integration wurde dabei von von Kotzebue (2014) näher untersucht, wodurch die beiden Komponenten *Informationsentnahmenähe und Konstruktionsnahe Integration* resultieren. Die Studie konnte zeigen, dass sich auch im Bereich der Integration die Informationsementnahme und die Konstruktion von Diagrammen unterscheiden. Die bereits durchgeführten Untersuchungen zur Integration beim Verstehen von Diagrammen (vgl. Lachmayer, 2008; von Kotzebue & Nerdel, 2012; von Kotzebue, 2014) bieten erste Anhaltspunkte zur Integration im Hinblick auf eine Darstellungsform. Um diesen Prozess genauer zu untersuchen, wurde zusätzlich auf das Strukturmodell der Visualisierungskompetenz von Wafi & Wirtz (2016) zurückgegriffen. Als zentrale Kategorie innerhalb des Modells steht das *Verknüpfen multipler Darstellungen*. Die Autoren betonen dabei zwei Teilaspekte, zum einen die Text-Bild-Integration als Verknüpfung von Text und Visualisierungen und die Bild-Bild-Integration, wenn mehrere Visualisierungen miteinander verknüpft werden sollen, d.h. der Fokus liegt auf so genannten Repräsentationswechseln. Dieser Repräsentationswechsel (vgl. Wafi & Wirtz, 2016) soll in der dritten Teilkomponente berücksichtigt werden. In Anlehnung an die Forderung zum Aufbau kohärenter Wissensstrukturen und eines konzeptionellen Verständnisses (vgl. Ainsworth, 1999; 2006; Seufert, 2003; Gehl, 2013) ist es bisher unerforscht, inwieweit die Komponente *Translation und Transformation* von mehr als zwei Repräsentationen einen Einfluss auf die Übersetzungsleistung hat. Gehl (2013) berichtet davon, dass globale Kohärenzbildungsprozesse Strukturwissen aufbauen können, wenn dabei eine hohe Vielfalt an Information in verschiedenen Modi verstanden werden muss. Dabei wäre es denkbar diese Prozesse ebenfalls in die Bereiche der *Informationsentnahmenähe* und *Konstruktionsnahen Integration* aufzuteilen. Im Rahmen dieser Untersuchung wird jedoch zunächst fokussiert, ob bei der Integration von mehr als zwei Repräsentationen ein empirischer Unterschied zu den beiden anderen Kompetenzkonstrukten *MERI* und *MERII* besteht.

Aus diesem Grund sind spezifische Kompetenzmodelle für diese Vielfalt schwierig zu klassifizieren, woraus ein hohes Forschungsinteresse resultiert. Dabei fokussieren empirische Studien Fragestellungen verschiedene Themenbereiche der Biologie, unterschiedliche Darstellungsformen oder Repräsentationstypen. Daraus folgt, dass domänenspezifische Kompetenzmodelle im Bereich der Kommunikationskompetenz (KMK, 2004) entweder übergeordnet für eine Fachdisziplin vorliegen (Kulgemeyer, 2010), sich auf eine spezifische Darstellungsform konzentrieren (Lachmayer, 2008; von Kotzebue, 2014) oder nur ausgewählte kognitive Prozesse der Text-Bild-Kombination untersuchen. Ein übergeordnetes Kompetenzmodell zum Prozess der Integration bei Text-Bild-Kombinationen, das gleichzeitig unterschiedliche Darstellungsformen sowie verschiedene Themengebiete der Biologie berücksichtigt und im Rahmen von Repräsentationskompetenz steht, findet sich aktuell nicht. Dabei sind gerade solche Kompetenzmodelle essentiell, da das Lernen mit Repräsentationen im Biologieunterricht so komplex organisiert ist wie es Unterschiede in den Darstellungsformen gibt.

3.3.2 DIMENSION AUFGABENKONTEXT

Der Begriff Kontext wird innerhalb empirischer Forschung sehr uneinheitlich definiert (van Oers, 1998, s. Kapitel 1.5). Eine genaue Einordnung des Kontextbegriffs solle an dieser Stelle nicht vorgenommen werden, es sei auf einschlägige Studien zur Kontextforschung verwiesen (vgl. u.a. Haugwitz, 2009; Bayrhuber et al., 2007; Gilbert, 2006; Finkelstein, 2005; Millar, 2005; Muckenfuß, 2004; Bennett, Hogarth, & Lubben, 2003). Ebenso die oft in diesem Zusammenhang stehende Interessensforschung zum kontextorientierten Lernen ist nicht Gegenstand dieser Untersuchung (vgl. u.a. Haugwitz, 2009; Millar, 2005; Benett et al., 2003; Krapp & Prenzel, 1992; 2011).

Im Rahmen der vorliegenden Arbeit werden unter Kontexten „Themen – oder thematische Aspektierungen eines fachlichen Teilgebiets“ (Muckenfuß, 2004, S. 64) verstanden. Van Vorst et al. (2014) sprechen dabei vom „Kontext als „storyline“ eines Problems“ und Leinhardt et al. (1990) ebenfalls von Problemsituationen, die mehr oder weniger kontextualisiert oder abstrakt sein können. Grundlage in der vorliegenden Studie sind die von Finkelstein (2005) klassifizierten Ebenen des Kontextbegriffs. Neben der äußeren Ebene, die außerschulischen Rahmenbedingungen und der zweiten Ebene, die Lernumgebung, wird in der innersten Ebene der Aufgabenkontext eingeordnet und im Rahmen dieser Arbeit berücksichtigt. Die Anforderung einer Aufgabe aus Kontextsicht betrachtet, ist vor dem Hintergrund dieser dritten Ebene auf die verwendeten fachlichen Inhalte innerhalb einer Testaufgabe beschränkt. Dabei wird auf die Kriterien von Gilbert (2006) Bezug genommen, wonach Kontexte der Vereinfachung

dienen sowie Konzepte verdeutlichen und damit schließlich Anwendungen ermöglichen sollen (vgl. Haugwitz, 2009).

Als Bezugsrahmen für den Aufgabenkontext dient das Modell von van Vorst et al. (2014). Dabei werden theoriegeleitet verschiedene Kategorien von Kontextmerkmalen zusammengetragen, die einer Kategorisierung von Kontexten dienen. So sollen Kontexte, Authentizität gewährleisten, hinreichend komplex sein, Bezüge zum Alltag herstellen, außergewöhnliche Phänomene behandeln und zusätzlich aktuelle Ereignisse fokussieren sowie die Relevanz behandelte Inhalte sicherstellen (vgl. van Vorst et al., 2014). Kulgemeyer (2009) spricht in diesem Zusammenhang von der Perspektive „adressatengerecht“, d.h. die gewählten Kontexte sind anschlussfähig und an den Interessen sowie dem Vorwissen der Lernenden orientiert. Gerade bei der Erfassung von Kompetenzen im Rahmen naturwissenschaftlicher Grundbildung geht es um die Bewältigung von realitätsnahen und alltagsrelevanten Kontexten (OECD, 2006), d.h. eine valide Messung steht im Vordergrund (Hartig & Jude, 2007). Schecker & Parchmann (2006) berichten im Rahmen von Entwicklung naturwissenschaftlicher Kompetenz, dass bisher nur wenig empirische Evidenz vorliegt, wenn es um die Antwort auf die Frage geht, ob die Problemlösefähigkeit vom jeweiligen Inhaltsbereich abhängt (vgl. Lind, Friege, & Sandmann, 2005; Lind & Friege, 2003). Ropohl, Walpuski, & Sumfleth (2014) berichten von Studien, worin die Komplexität des Inhalts einen Einfluss auf die Aufgabenschwierigkeit hat (vgl. Bernholt, 2010; Kauertz, 2008; Kauertz et al., 2010). Watzka & Girwidz (2015) konnten hingegen keinen Einfluss auf die Schwierigkeit diagnostizieren.

Aus diesem Grund wurde in das Kompetenzstrukturmodell die Komponente *Aufgabenkontext* aufgenommen, damit Repräsentationskompetenz in verschiedenen Teilgebieten innerhalb der Domäne Biologie untersucht werden kann. Die Wahl auf die drei Kontexte Ökologie, Stoffwechsel und Genetik wird damit begründet, dass Studien zu Abituraufgaben zeigen, dass gerade diese drei Themenbereiche eine große Rolle in der Wahl und den Anforderungen im Abitur spielen (vgl. Florian, 2012, s. a. Kapitel 1.5 und Kapitel 4.2). Zudem zeigen diese Inhalte eine hohe curriculare Relevanz in Schule und Hochschule. Damit leistet die vorliegende Arbeit einen Beitrag zum Vergleich von Repräsentationskompetenz innerhalb ökologisch valide ausgewählter Inhaltsbereiche und Anforderungen in der Biologie und auf die Frage, ob repräsentationsspezifische Kompetenzen inhaltsabhängig sind oder nicht.

3.3.3 DIMENSION REPRÄSENTATIONSTYP

Innerhalb der Sichtstruktur von Aufgaben kann u.a. das Aufgabenmerkmal Darstellungsform berücksichtigt werden, das die Darstellung von Informationen in unterschiedlichen Reprä-

sentationen beschreibt (Leisen, 2005). Prenzel et al. (2002) berichten, dass die Darstellungsform eine Auswirkung auf die Schwierigkeit von Testaufgaben haben kann. Aus diesem Grund wird die Dimension *Repräsentationstyp* in das Kompetenzstrukturmodell aufgenommen, um zwei unterschiedliche Formen externer Repräsentationen zu fokussieren. Das Modell konzentriert sich dabei auf die Darstellungsformen *Diagramm* und *Schemazeichnung*, worunter Zeichnungen unterschiedlicher Abstraktion und Realitätsnähe berücksichtigt werden (s. Kapitel 1.4.2). Im Hinblick auf Diagramme liefert das Modell einen Zugewinn, indem nicht nur ein spezifischer Fachinhalt untersucht wird, sondern Diagramme in unterschiedlichen Aufgabenkontexte innerhalb der Domäne Biologie fokussiert werden. Aufgrund der theoretisch fundierten Annahme unterschiedlicher Formen von Schemazeichnungen kann ein Beitrag zur Klassifikation und genaueren Abgrenzung realistischer und logischer Bilder erreicht werden. Die Befunde sind stark heterogen, so können Schemazeichnungen den logischen Bildern zugeordnet werden, wenn die Konventionen für das Verstehen und Lösen bekannt sein müssen (vgl. Griffard, 2013; Ainsworth, 2003; Kosslyn, 1989). Sind diese nicht darin einordbar, können in Schemazeichnungen realistische bis hin zu ikonisch, domänenspezifisch auf Konventionen festgelegt Elemente vorkommen (vgl. Weidenmann, 2002). Derartige Schemazeichnungen werden schließlich als semirealistische Darstellungen bezeichnet, die eine ansteigende Abstraktion aufweisen (Griffard, 2013; Tsui & Treagust, 2013). In diesem Zusammenhang leistet die vorliegende Arbeit einen Beitrag zur weiterführenden Forschung und Charakterisierung von Schemazeichnungen, indem Merkmale unterschieden und anschließend klassifiziert werden können. Fokussiert werden dabei Schemazeichnungen, die zeigen, wie etwas funktioniert oder organisiert ist (vgl. Griffard, 2013). Gleichzeitig kann anhand der unterschiedlichen Aufgabenkontexte Aufklärungsarbeit geleistet werden, indem konventionalisiertes, domänenspezifisches Vorwissen in den unterschiedlichen biologischen Schemata untersucht wird.

3.4 ZUSAMMENFASSUNG

Das Bildungswesen steht vor der ständigen Herausforderung aus Weiterentwicklung und Optimierung von Bildungsprozessen (Hartig & Jude, 2007). Die Outcome-Orientierung führte zur Einführung nationaler Bildungsstandards und zur Entwicklung von Kompetenzmodellen (Bernholt et al., 2009). Aus Sicht der Fachdidaktiken bieten Standards eine Orientierungsfunktion, indem sie „erreichbare Teilziele als Bestandteile übergeordneter, normativ und empirisch begründeter Zieldimensionen ausweisen“ (Schecker & Parchmann, 2005, S. 48). Kompetenzen enthalten die Merkmale Erlernbarkeit, Kontextspezifität und Anforderungen in den fachbezogenen Leistungsbereichen (Hartig & Klieme, 2006). Die in den nationalen Bildungsstandards verankerten kommunikativen Kompetenzen, das Anwenden von idealtypischen Darstellungsformen auf komplexe Sachverhalte, verlangt neben dem Erschließen von biologischen Informationen auch das selbstständige Konstruieren und Integrieren. Neben einzelnen Repräsentationen sind in diesem Zusammenhang auch MER von entscheidender Bedeutung, im Hinblick auf Integrationsprozesse bisher aber wenig untersucht.

Das vorliegende postulierte Kompetenzstrukturmodell zum Umgang mit MER berücksichtigt dabei biologiespezifische Repräsentationen als auch die Integration in unterschiedlichen Aufgabenkontexten und bei verschiedenen Bildtypen. Entsprechend kann das Modell zu einer erhöhten Kohärenz zwischen Curriculum, fachspezifischen Herausforderungen sowie kognitionspsychologischen Modellen und konkreter Aufgabensammlung führen (vgl. Klieme & Leutner, 2006). Damit verbunden hat das Kompetenzmodell die Chance einen Beitrag zur Qualitätssicherung und Qualitätsentwicklung in Bildungssystemen zu leisten (Bernholt et al., 2009). Aus diesem Grund stellt die theoriegeleitete Entwicklung des vorgestellten Kompetenzstrukturmodells eine weiterführende Entwicklungsarbeit in der biologiedidaktischen Forschung dar.

III ZIELE, FORSCHUNGSFRAGEN UND HYPOTHESEN

Unterschiedliche Forschungsarbeiten fokussieren verschiedene Facetten beim Umgang mit multiplen externen Repräsentationen (MER). Weder theoretische Arbeiten noch empirische Studien zeigen ein einheitliches Bild im Hinblick auf Repräsentationskompetenz für den Biologieunterricht. Ziel dieser Arbeit ist die Entwicklung und empirische Überprüfung eines biologiepädagogischen Kompetenzstrukturmodells zum Umgang mit MER. Im Fokus steht dabei die Entwicklung und Modellierung eines Kompetenzmodells, das unterschiedliche Integrationsprozesse bei MER sowie unterschiedliche Aufgabenkontexte und Repräsentationstypen berücksichtigt.

Darauf basierende, ergeben sich für die vorliegende Arbeit vier zentrale Teilziele.

- I. Überprüfung des postulierten Kompetenzmodells auf seine Gültigkeit durch Anwendung von qualitativen und quantitativen Testverfahren.
- II. Diagnose von Kompetenzen bei der Integration von multiplen externen Repräsentationen (MER), die eine *Informationsentnahmenähe Integration (MERI)*, *Konstruktionsnähe Integration (MERII)* und *Translation oder Transformation aus mehreren Repräsentation (MERIII)* erfordern.
- III. Ermittlung der Aufgabenschwierigkeit, die durch Merkmale aus den Dimensionen im Modell (*Integration von MER, Aufgabenkontext* und *Repräsentationstyp*) bestimmt werden und zur Identifikation schwierigkeiterzeugender Aufgabenmerkmale herangezogen werden kann.
- IV. Aufklärung von Zusammenhängen zwischen Personenmerkmalen und Repräsentationskompetenz.

Daraus leiten sich folgende Forschungsfragen und Hypothesen ab.

Inwiefern bildet sich die theoretisch abgeleitete Struktur des Kompetenzmodells in den empirischen Daten ab?

Auf Basis von kognitionspsychologischen Grundlagen und empirischen Befunden werden die Kompetenzen zum Umgang mit multiplen externen Repräsentationen in drei Teilkomponenten differenziert (s. Kapitel 3.3.1).

Dabei geht es um Text-Bild-Kombinationen, die komplementäre Informationen enthalten (z.B. Saß, 2012, Cook, 2006) und eine explizite Integrationsleistung als Prozess für ein vertieftes Verständnis erfordern (Cook, 2006). Aus diesem Grund wird neben der Interpretation als *Informationsentnahmenahe Integration (MERI)*, der Konstruktion als *Konstruktionsnahe Integration (MERII)* auch die Teilkomponente *Translation von mehr als zwei Repräsentationen* und *Transformation mit einer weiteren bildlichen Repräsentation als Output (MERIII)* unterschieden.

Bevor geprüft werden kann, ob der Kompetenztest repräsentationsspezifisches Wissen beim Umgang mit MER misst, wird in einem ersten Schritt die Modellgültigkeit hinterfragt. Die interne Konsistenz gilt als Maß für die Homogenität einer Skala, d.h. alle zugrundeliegenden Testitems sollten ein gemeinsames latentes Konstrukt messen (Wellnitz, 2012). Durch die Prüfung der Reliabilitäten der geschätzten Skalen kann der Anteil der Varianz beschrieben werden, der tatsächlich auf Personenunterschiede zurückzuführen ist (Rost, 2004). Im Anschluss daran geht es um die empirische Überprüfung der postulierten Modellstruktur mit den ermittelten Daten. Die Überprüfung erfolgt über den Vergleich der Modellpassung anhand verschiedener Maße (s Kapitel 8.2). Dabei wird der Frage nachgegangen, wie viele unterschiedliche Kompetenzkonstrukte sich bei der Integration von MER identifizieren lassen oder ob die Fähigkeiten der Studierenden zum Umgang mit MER in einem eindimensionalen Modell durch eine übergeordnete *Repräsentationskompetenz* abgebildet werden. Auch bei empirischer Abgrenzbarkeit mehrerer Teilkompetenzen innerhalb einer Dimension sind Korrelationen zwischen diesen Teilkomponenten zu erwarten, da sie Teil eines übergeordneten Konstrukts sind. Untersuchungen zu Modellkompetenz, Erkenntnisgewinnung, Diagrammkompetenz und Biologische Informationen aus unterschiedlichen Repräsentationen schließen, bestätigen meist mittlere bis hohe Korrelationen zwischen den differenzierten Teilkomponenten (vgl. Terzer, 2012; Wellnitz, 2012, Lachmayer, 2008; Ziepprecht, 2016). Vor diesem Hintergrund sind derartige Korrelationen auch bei Vorliegen der postulierten mehrdimensionalen Struktur nicht auszuschließen.

Entsprechend werden folgende Hypothesen aufgestellt:

H_{1.0}: Die Dimension *Integration von MER* lässt sich durch eine Dimension abbilden. Mehrdimensionale Modelle führen zu keiner besseren Passung als das eindimensionale Modell ($p < 0,05$).

H_{1.1}: Die Dimension *Integration von MER* lässt sich durch die drei Teilkomponenten *Informationsentnahmenähe Integration (MERI)*, *Konstruktionsnähe Integration (MERII)* und *Translation und Transformation (MERIII)* beschreiben, d.h. empirisch erfolgt die beste Passung bei einem dreidimensionalen Modell.

Der Umgang mit MER und die damit verbundene Übersetzungsleistung ist in den biologiespezifischen Teilgebieten mit unterschiedlichen Abstraktionsgraden von Bildern und verschiedenen zugrundeliegenden Konventionen verbunden (Roth & Pozzer-Ardenghi, 2013; Griffard, 2013; s. Kapitel 1.4.2). Ein Aspekt, der dabei bisher nicht erfasst wurde, ist der Einfluss des biologiespezifischen Kontexts auf die Repräsentationsleistung sowie mögliche Unterschiede zwischen verschiedenen Darstellungsformen. In diesem Zusammenhang stehen die Konstrukte *Aufgabenkontext* und *Repräsentationstyp* (s. Kapitel 3.3.2 und 3.3.3). Für eine curricular möglichst breite Verankerung biologiespezifischer Kontexte werden die Teilgebiete Ökologie, Stoffwechsel und Genetik ausgewählt (s. Kapitel 1.5). Als Darstellungsformen werden einerseits Diagramme berücksichtigt, da sie an bisherige Studien anknüpfen (z.B. Lachmayer, 2008; von Kotzebue, 2014) und andererseits Schemazeichnungen, um neue Erkenntnisse in Zusammenhang mit Repräsentationskompetenz zu gewinnen. Vor diesem Hintergrund werden die Konstrukte *Aufgabenkontext* und *Repräsentationstyp* als weitere latente Variablen definiert und folgende Hypothesen formuliert:

H_{1.2}: Die Dimension *Aufgabenkontext* lässt sich durch die drei Teilkomponenten *Ökologie*, *Stoffwechsel* und *Genetik* beschreiben, d.h. empirisch liegt ein dreidimensionales latentes Konstrukt vor.

H_{1.3}: Die Dimension *Repräsentationstyp* lässt sich durch die zwei Teilkomponenten *Diagramm* und *Schemazeichnung* beschreiben, d.h. empirisch liegt ein zweidimensionales latentes Konstrukt vor.

Welche schwierigkeitsgenerierenden Merkmale können in den Aufgaben zum Umgang mit multiplen externen Repräsentationen identifiziert werden?

Auf Basis einer inhaltsstrukturbezogenen Aufgabenanalyse zum Umgang mit MER im biologischen Kontext, erfolgt die Ermittlung der schwierigkeitsgenerierenden Merkmale unter dem Inhaltsaspekt (vgl. Kauertz, 2008). Bisher weiß man wenig darüber, welche Faktoren die Übersetzungsleistung bei Text-Bild-Kombinationen beeinflussen. In diesem Zusammenhang geht es um unterschiedliche kognitive Prozesse zur Integration von MER vor dem Hintergrund biologiespezifischer Inhalte und der Verwendung unterschiedlicher Repräsentationstypen, die domänenspezifisches Wissen über Konventionen erfordern. Vergleichbar zur Modellgeltung wird angenommen, dass die unterschiedlichen Konstrukte auch die Aufgabenschwierigkeit beeinflussen. Entsprechend werden folgende Hypothesen formuliert:

H_{2.1}: Die Schwierigkeit der Testitems beruht auf der Dimension *Integration von MER* mit den Teilkomponenten MERI, MERII, MERIII.

H_{2.2}: Die Dimension *Aufgabenkontext* mit den Teilkomponenten Ökologie, Stoffwechsel und Genetik hat einen Einfluss auf die Itemschwierigkeit.

H_{2.3}: Die Dimension *Repräsentationstyp* mit den Teilkomponenten Ökologie, Stoffwechsel und Genetik hat einen Einfluss auf die Itemschwierigkeit.

Welche Zusammenhänge bestehen zwischen Personenfähigkeiten und Kompetenz und welche Kompetenzausprägungen können damit diagnostiziert werden?

Die Ausprägung der Kompetenzen von Lernenden zum Umgang mit MER wird, abhängig vom bestätigten Modell, für eine oder mehrere Dimensionen vorgenommen. Studien zeigen, dass Text-Bild-Kombinationen zu besseren Behaltens- und Verstehensleistungen bei Lernenden führen können (z.B. Hegarty & Just, 1993; s. Kapitel 1.4.3 und Kapitel 2.4). Daneben wird ebenfalls berichtet, dass es beim Lernen mit MER zu Schwierigkeiten und Problemen kommen kann (z.B. Ainsworth, 2006; Kozma, 2003; Kozma & Russell, 1997; s. Kapitel 1.4.3). Entscheidend dabei sind das Vorliegen unterschiedlich abstrakter Darstellungsformen und die anschließend geforderte Übersetzungsleistung. Vor diesem Hintergrund gelingt es Lernenden mit hohem Vorwissen besser, die relevanten Oberflächenkriterien zu erschließen als Novizen (z.B. Cook, Wiebe, & Carter, 2008; Kozma & Russell, 1997). Im Rahmen der Diagrammkompetenz und zum thematischen Kontext der Fotosynthese scheint die Konstruktion insgesamt leichter für Lernende zu sein als die Informationsentnahme (vgl. Lachmayer, 2008;

von Kotzebue, 2014). Dies wird als Hinweis darauf gewertet, dass die Integration von MER je nach Teilkomponente unterschiedlich schwierig sein kann. Allerdings zeigen verschiedene Studien auch, dass repräsentationsspezifisches Wissen stark von der Art der Bilder und ihrer zugrundeliegenden Konvention (z.B. Brandstetter-Korinth, 2017; Ziepprecht, 2016) oder dem Aufgabenkontext abhängen kann (z.B. Werner, Schwanewedel, & Mayer, 2014). Dieser Zusammenhang ist vor dem Hintergrund des integrativen Verhältnisses zwischen Text und Bild in der Biologiedidaktik bisher wenig untersucht, entsprechend werden die folgenden Hypothesen aufgestellt:

H_{3.1}: Unter Voraussetzung, dass ein mehrdimensionales Modell empirisch bestätigt werden kann, weisen die Leistungen der Studierenden bei der *Integration von MER* in den Teilkomponenten *MERI*, *MERII*, *MERIII* Unterschiede auf.

H_{3.2}: Unter Voraussetzung, dass ein mehrdimensionales Modell empirisch bestätigt werden kann, weisen die Leistungen der Studierenden bei der Dimension *Aufgabenkontext* in den Teilkomponenten *Ökologie*, *Stoffwechsel* und *Genetik* Unterschiede auf.

H_{3.3}: Unter Voraussetzung, dass ein mehrdimensionales Modell empirisch bestätigt werden kann, weisen die Leistungen der Studierenden bei der Dimension *Repräsentationstyp* in den Teilkomponenten *Diagramm* und *Schemazeichnung* Unterschiede auf.

Welche Zusammenhänge bestehen zwischen Personenmerkmalen und Kompetenz?

Für das in der vorliegenden Arbeit untersuchte Konstrukt, das repräsentationsspezifisches Wissen in unterschiedlichen Inhalten der Biologie untersucht, kann ein gewisser Zusammenhang zwischen Studiengang, den Schulnoten in Biologie und Mathematik und der Kompetenz angenommen werden. Hierzu liegen im hochschuldidaktischen Bereich keine einschlägigen Studien vor. Verschiedene biologiedidaktische Untersuchungen mit Schülerinnen und Schülern als Stichprobe, berichten von meist geringen Korrelationen oder überhaupt keinem signifikanten Zusammenhang (z.B. Wellnitz, 2012; Grube, 2011; Eggert & Bögeholz, 2010; Schmiemann, 2010). Aus diesem Grund werden Zusammenhänge zwischen den vorliegenden Personenmerkmalen explorativ untersucht. Zusätzlich gilt es das Geschlecht zu kontrollieren, damit in Bezug auf das Repräsentationswissen angehender Studierender keine geschlechtsabhängigen Unterschiede diagnostiziert werden. Vor diesem Hintergrund werden folgende Hypothesen formuliert:

H4.1: Es bestehen Kompetenzunterschiede zwischen Personen verschiedener Studiengänge.

H4.2: Die Kompetenzen der Studierenden korrelieren mit der Biologienote.

H4.3: Die Kompetenzen der Studierenden korrelieren mit der Mathematiknote.

H4.4: Es bestehen keine Kompetenzunterschiede zwischen weiblichen und männlichen Studierenden.

IV EMPIRISCHER TEIL

4 MATERIAL DER VORSTUDIE

Zur empirischen Überprüfung des Kompetenzstrukturmodells (s. ABBILDUNG 7) wurden zwei Studien durchgeführt, die Pilotierung und die Hauptstudie der Arbeit. Das zentrale Ziel der Vorstudie ist die Erstellung eines Kategoriensystems zur Kodierung offener Testaufgaben sowie die Überprüfung der modellkonform konstruierten Testitems auf ihre Verträglichkeit mit dem angewendeten Testmodell. Zusätzlich wird eine erste Rasch-Analyse durchgeführt und untersucht, ob sich die angenommenen Dimensionen des Kompetenzstrukturmodells *Integration multipler externer Repräsentationen (MER)* in weitere Subdimensionen ausdifferenzieren lassen.

4.1 STICHPROBE

Die Population der vorliegenden Untersuchung waren Studierende des ersten und zweiten Semesters an unterschiedlichen Universitäten. Da die Untersuchung einen curricularen und bildungspolitischen Bezug aufweist, wurde sichergestellt, dass sich unter den Versuchspersonen vor allem Lehramtsstudierende unterschiedlicher Fachrichtungen befinden. Die Stichprobe für die Pilotierung setzte sich aus 730 Studierenden ($N_{A\text{-Serie}}=394$, $N_{B\text{-Serie}}=336$) unterschiedlicher Studienfachrichtungen zusammen. Beteiligte Universitäten waren die Technische Universität München (TUM), die Ludwig-Maximilians-Universität München (LMU), die Universität Regensburg sowie die Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Die Verteilung der Studierenden, die an der Pilotierung teilgenommen haben, sowie die jeweilige Studienfachrichtung sind in TABELLE 1 dargestellt.

TABELLE 1: VORSTUDIE - STUDIENGÄNGE UND ANZAHL DER STUDIERENDEN (N=730)

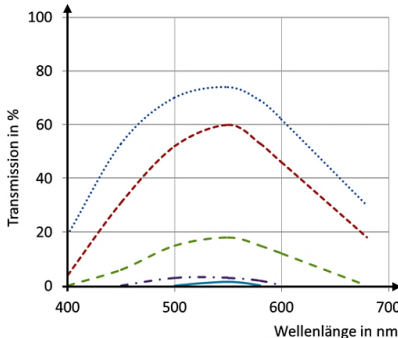
Studiengang	Anzahl der Studierenden
Lehramt Gymnasium und Realschule Fächerkombination: Biologie/ Chemie, Mathematik/ Chemie, Geographie/ Chemie	64
Lehramt Grund-, Mittel- und Förderschule	28
Maschinenwesen und TUM-BWL	26
Mathematik und Informatik	56
Sportwissenschaften	91
Ernährungswissenschaften	65
Brauwesen und Getränketechnologie	80
Biowissenschaften	41
Molekulare Biotechnologie	55
Bioprozesstechnik	67
Lebensmitteltechnologie	57
Agrar- und Gartenbauwissenschaften	95
Sonstige oder keine Angaben	5

4.2 TESTKONSTRUKTION

Zur empirischen Überprüfung des Kompetenzstrukturmodells (s. ABBILDUNG 7) wurde ein Testinstrument konstruiert, welches die Erfassung der durch das Modell formulierten Dimensionen und deren Aufteilung durch Teilkomponenten realisiert. Das postulierte theoretische Kompetenzstrukturmodell wird dadurch in ein messbares psychometrisches Modell überführt (Mayer & Wellnitz, 2014).

Bei der Gestaltung der Testhefte wurde auf eine einheitliche Struktur geachtet. Nach der Titelseite werden in der Einleitung die Durchführung, die vorgesehene Zeit für die Testung sowie Hinweise zum Bundesdatenschutz nach § 4 beschrieben. Als Hintergrundvariablen, die aufgrund der Literaturhinweise auf einen möglichen Einfluss auf die Testleistung der Studierenden schließen können, wurde der *Lernerhintergrund* erfragt sowie *Lernermerkmale* erhoben. Hierzu folgte nach der Einleitung ein Kurzfragebogen zur Person. Dieser diente zur

Erfassung des Lernerhintergrunds, d.h. der soziodemographischen Daten wie dem Geschlecht, dem Alter und dem Bildungshintergrund der Studierenden. Dabei wurde das Bundesland des Erwerbs der Allgemeinen Hochschulreife erfragt, bis zu welcher Jahrgangsstufe Biologie in der Schule besucht wurde und ob und mit welcher Leistung das Abitur in den Fächern Biologie oder Mathematik abgelegt wurde. Die Leistung wurde dabei im gängigen Punktesystem für Noten in der Oberstufe (Punkte von 0 bis 15) abgefragt. Zusätzlich wurden die Informationen zu aktuellem Studiengang, dem laufenden Fachsemester und der Universität erhoben. Im Rahmen der Vorstudie wurden zwei zusätzliche Erhebungsinstrumente mit in die Untersuchung aufgenommen, die Aussagen zu den Lernermerkmalen liefern sollen. Im Anschluss an den Kurzfragebogen zur Person und vor den Testaufgaben folgte ein Kurzfragebogen zur Diagrammkompetenz. Dieser diente zur Erfassung von diagrammspezifischem Vorwissen hinsichtlich der allgemeinen Diagrammkonventionen und Gestaltungsprinzipien. Ein Beispiel für zwei Items aus diesem Kurzfragebogen zeigt die folgende ABBILDUNG 8.



1. Welchem der folgenden Diagrammtypen entspricht das oben dargestellte Diagramm?

<input type="checkbox"/> Kreisdiagramm	<input type="checkbox"/> Liniendiagramm
<input type="checkbox"/> Säulendiagramm	<input type="checkbox"/> Balkendiagramm

2. Zuordnung der Variablen:

a) Auf der x-Achse steht die abhängige Variable

richtig

falsch

Die x-Achse wird auch als _____ bezeichnet.

ABBILDUNG 8: BEISPIEL FÜR ZWEI ITEMS ZUM DIAGRAMMSPEIFISCHEN VORWISSEN

Im Anschluss an diesen diagrammspezifischen Kurzfragebogen folgen die eigentlichen Testaufgaben. Am Ende eines jeden Testhefts der Vorstudie folgte ein C-Test zur Einschätzung der Lesekompetenz. Hierfür wurde der Text des Aufgabenbeispiels „Vitamine“ und „Semesterferien“ aus dem Aufnahmetest Deutsch vom Studienkolleg Sachsen der Universität Leipzig (2006) herangezogen. Die Aufgabentexte sind relativ kurz und prüfen die Textproduktion im Zusammenhang ab.

Die Konstruktion eines geeigneten Leistungstests (Definition vgl. Leutner et al., 2008) ist Voraussetzung, wenn von der tatsächlich beobachteten Fähigkeit auf eine relevante Kompetenz geschlossen werden soll (Klieme, Hartig, & Rauch, 2008). Leistungstests sind aus Skalen zusammengesetzt, die eine Facette des relevanten Konstrukts messen (Leutner et al., 2008). In der vorliegenden Arbeit ist die Skala damit eine der drei Komponenten aus dem Kompetenzstrukturmodell wie die *Integration multipler externer Repräsentationen (MER)* und die Facette wiederum z.B. die *Informationsentnahmenehe Integration (MERI)*. Als Aufgabe wird dabei „eine inhaltlich zusammenhängende, nicht aufteilbare Einheit, in der ein Problem oder eine bestimmte Einheit von Material (z.B. ein Lesetext) vorgegeben wird“ (Hartig & Jude, 2007, S. 19) angesehen. Im Unterschied dazu wird ein Item als kleinste Analyseeinheit eines Tests definiert, und wiederum aus einem Aufgabenstamm und einem Antwortformat besteht (Hartig & Jude, 2007; Rost, 2004).

Die Schritte der Testentwicklung im Anschluss an die theoriegeleitete Entwicklungsarbeit und die messmethodischen Überlegungen können zusammenfassend dargestellt werden. Der Prozess ist dabei an die Untersuchungen von Terzer et al. (2013) sowie Mayer & Wellnitz (2014) angelehnt und in ABBILDUNG 9 dargestellt.

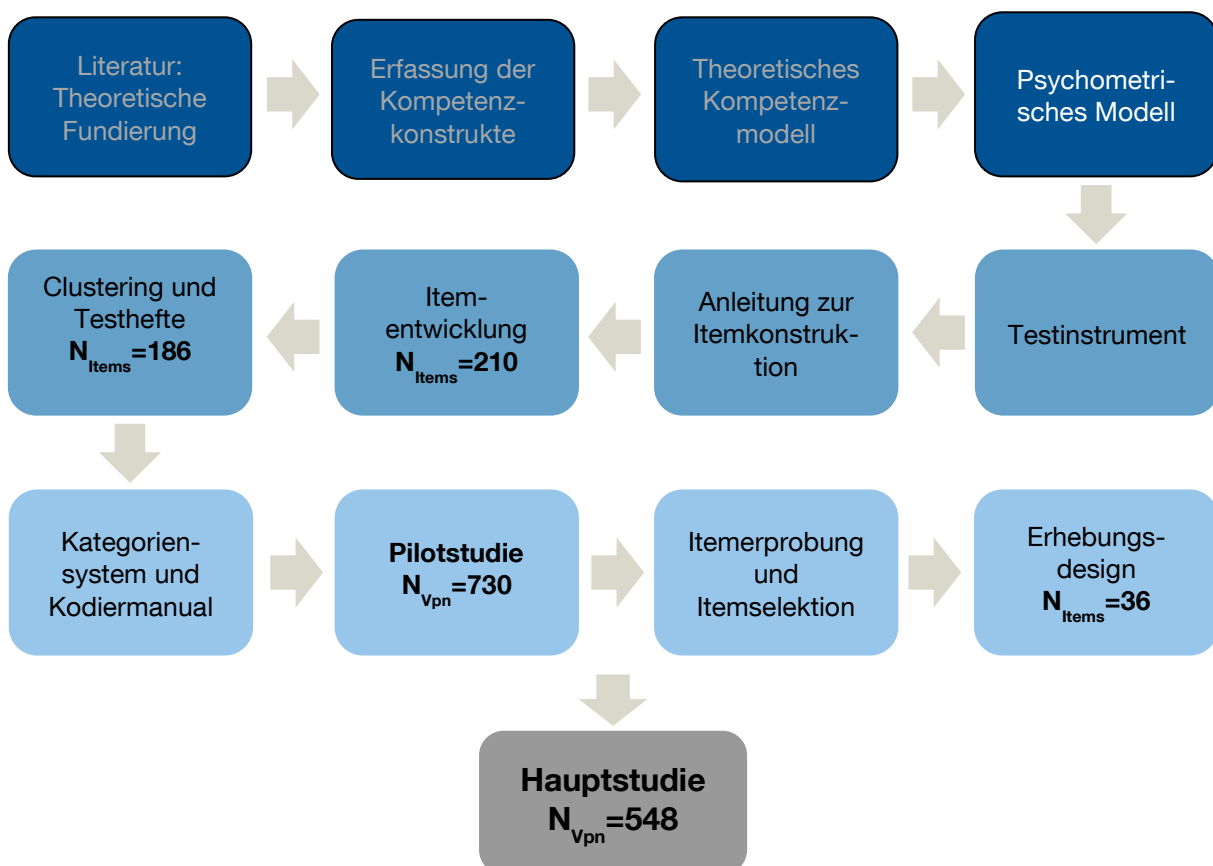


ABBILDUNG 9: SCHRITTE DER TEST- UND ITEMKONSTRUKTION UND DURCHFÜHRUNG DER VOR- UND HAUPTSTUDIE

Die Entwicklung aller Testaufgaben erfolgte auf Grundlage des in Kapitel 3.3 vorgestellten Kompetenzstrukturmodells zum *Umgang mit multiplen externen Repräsentationen (MER) im biologischen Kontext*. Um das entwickelte Kompetenzmodell enger an den Unterricht anbinden zu können, ist eine transparente Testaufgabenkonstruktion und damit Operationalisierung entscheidend (Bernholt, Parchmann & Commons, 2009). Für eine effiziente Entwicklung eines Tests zum Umgang mit multiplen externen Repräsentation im biologischen Kontext wurden die sieben Schritte der Test- und Itemkonstruktion auf der Grundlage von Kompetenzmodellen herangezogen (vgl. Terzer et al., 2013). Hierzu gehört neben der Formulierung der theoretischen Fundierung die Festlegung der Testkonzeption, d.h. der Anwendungskontext und die Testart werden definiert und daraufhin die Itemkonstruktion systematisiert. Anschließend folgt auf Basis von Indikatoren, dem Itemuniversum und dem Antwortformat die Entwicklung einer Konstruktionsanleitung. Diese gewährleistet eine Passung der Testaufgaben zum entwickelten Kompetenzstrukturmodell (Köller et al., 2008a). In der Itemkonstruktionsanleitung wurden der Aufgabenstamm definiert, standardisierte Aufgabenstellungen mit Operatoren formuliert und mögliche Antworten, formuliert als Erwartungshorizont oder Ankreuzen der richtigen Lösung, beschrieben. Damit kann eine transparente modellbasierte Entwicklung von Testitems erfolgen, die einerseits die theoretisch abgeleiteten Strukturen innerhalb des Modells vollständig repräsentieren und andererseits zu einem ökonomischen Test zusammengestellt werden können. Im Anschluss an die Itementwicklung folgt die Itemprobung und –selektion sowie die Festlegung des Erhebungsdesigns, abschließend wird der Test für den intendierten Anwendungskontext eingesetzt (vgl. Terzer et al., 2013). Damit kann schließlich ein Beitrag zur Qualitätssicherung geleistet und eine valide Interpretation der Testwerte ermöglicht werden (Haladyna, 2004).

4.2.1 OPERATIONALISIERUNG

Zentrales Element der Operationalisierung ist die Verknüpfung der Aufgaben mit dem theoretischen Konstrukt (Kauertz, 2014). Ein Kompetenztest ist jedoch „[...] viewed as a sample of tasks from this large domain“ (Shavelson, 2013, S. 40). Testaufgaben beschreiben relevante Situationseigenschaften als Teile des Kompetenzkonstrukts und in Form von schwierigkeiterzeugenden Aufgabenmerkmalen (Schmiemann, 2010). Damit berücksichtigen Testaufgaben immer nur einen Teil von Situationen aus dem zu untersuchenden Kompetenzbereich, die empirisch erfasst werden können. Im Rahmen der vorliegenden Untersuchung wird der Umgang mit multiplen externen Repräsentationen durch das Verhältnis von vorgegebenen Informationen im Aufgaben- und Itemstamm und der erwarteten Integrationsleistung MERI, MERII, MERIII operationalisiert (Kauertz et al., 2010). Die Dimension MER wird dabei

durch gezielte Variation mit der Dimension Aufgabenkontext und der Dimension Repräsentationstyp in Beziehung gesetzt. Jede modellkonforme konstruierte Testaufgabe kann eindeutig auf jeder der drei Dimensionen verortet werden. Die Testaufgaben wurden so konstruiert, dass keine redundanten Informationen in den Aufgaben enthalten sind, d.h. die gleichen Informationen wurden nicht in zwei unterschiedlichen Repräsentationstypen wie z.B. Text und Bild und damit doppelt präsentiert. Studien zeigen, dass die Text-Bild-Integration bei redundanten Informationen beeinträchtigt werden kann (Sweller & Chandler, 1994; vgl. Cook, 2006). Daneben ist eine fachliche Passung der Items entscheidend (Webb, 2006). D.h. als Grundvoraussetzung für einen validen Test gilt, dass die Testaufgaben zu den fachlichen Themengebieten passen und diese Inhalte adäquat mit typischen Repräsentationen abbilden. Damit enthält der Kompetenztest Aufgaben zu unterschiedlichen biologischen Themen der Sekundarstufe II aus den Aufgabenkontexten Ökologie, Stoffwechsel und Genetik, die ferner zwischen den beiden Bildtypen Diagramm und Schemazeichnung variieren. Vor diesem Hintergrund können schwierigkeiterzeugende Merkmale identifiziert werden. So unterscheidet beispielsweise Saß (2012) drei wesentliche Kriterien, wenn es um die erfolgreiche Beantwortung einer Testaufgabe geht: a) die Gestaltung der Aufgabe, d.h. die Kontextualisierung und das Antwortformat, b) die kognitive Anforderung der Aufgaben, d.h. die MER-Integrationsleistung und c) die Personenmerkmale.

4.2.2 AUFGABENKONSTRUKTION

Im Rahmen der vorliegenden Untersuchung wurden domänenspezifische und kontextgebundene Testaufgaben konstruiert, die sich für die Befragung einer breiten Zielgruppe im Anschluss an das Abitur eignen. Dabei sollen die Testaufgaben über einen großen Schwierigkeitsbereich streuen und inhaltlich in den drei genannten Aufgabenkontexten breit gefächert sein. Dadurch wird versucht möglichst viele Merkmalsausprägungen abzudecken und empirisch zu erfassen. Die Testaufgaben wurden so entwickelt, dass sie ausschließlich Informationen, die im direkten Zusammenhang mit der Bearbeitung der unterschiedlichen Konstrukte aus dem entwickelten Kompetenzmodell stehen. Bei der Aufgabenkonstruktion wurde darauf geachtet, dass nur diese Fähigkeiten für die Bearbeitung relevant sind und keine anderen Konstrukte abgefragt werden. Kauertz (2014) schlägt folgende Bedingungen bei der Aufgabenentwicklung für einen Rasch-skalierten Test vor:

- Die Aufgabe muss das theoretische Konstrukt möglichst eindeutig und ausschließlich repräsentieren.
- Die Aufgabe muss Gemeinsamkeiten zu anderen Aufgaben im Test zeigen.

- Die Aufgabe muss mindestens ein definiertes schwierigkeitsgenerierendes Merkmal aufweisen.

Diese Bedingungen werden als Grundlage der Aufgabenentwicklung definiert. Neben einer modellbezogenen Neuentwicklung von Testaufgaben, wurden auch vorhandene Aufgaben in Schul- und Lehrbüchern herangezogen, überarbeitet und in das zugrundeliegende Kompetenzmodell eingeordnet (vgl. u.a. Theyßen et al., 2006). Damit ein möglichst großer Anteil der Varianz in den Antworten der Studierenden sowie der Entwicklungsprozess der Items nachvollziehbar und theoriegeleitet erklärt werden kann, wurden die Itemmerkmale in einer Konstruktionsanleitung festgehalten (vgl. Kauertz, 2008; Terzer et al., 2013). Aus Gründen der Validierung wurde neben einer objektiven Konstruktionsanleitung zusätzlich eine Expertenbeurteilung der Aufgaben hinsichtlich der Passung zum Konstrukt durchgeführt (vgl. Kauertz, 2014). Vor diesem Hintergrund konnte eine Übereinstimmung der Zuordnung zu den drei Teilkomponenten der MER-Integration mit über 90% erreicht werden (s. Kapitel 6.1). Alle nicht stimmigen Zuordnungen wurden diskutiert und abschließend übereinstimmend klassifiziert. Im Anschluss daran wurden die Aufgabenantworten gezielt in die Zellen der Kompetenzmatrix, bestehend aus den drei unterschiedlichen MER-Integrationen, Repräsentationstyp und Aufgabenkontext, eingeordnet. Vor diesem Hintergrund kann vom Testverhalten auf die individuelle Ausprägung des latenten Konstrukts geschlossen werden (Hartig, Frey, & Jude, 2012). Folgende Elemente wurden in die finale Konstruktionsanleitung aufgenommen (vgl. Kauertz, 2008; Neumann, 2011; Terzer et al., 2013):

- Beschreibung der Kompetenz, die mit dem Item erfasst werden soll.
- Quellenangabe
- Aufgabentitel mit Bezug zum fachlichen Hintergrund
- Beschreibung der Informationen, die im Aufgabenstamm enthalten sein sollen
- einleitender Text als Impuls
- fachliche Beschreibung der Inhalte sowie Fachbegriffe, die zur Beantwortung benötigt werden
- Formulierung eines standardisierten Aufgabenimpulses für jedes Element der Theorie, zu dem Items entwickelt wurden
- Beschreibung der Informationen, die im Itemstamm enthalten sein sollen und Fragestellung als Handlungsaufforderung über Operatoren
- Beschreibung der Antwortmöglichkeiten in Form eines Erwartungshorizontes

Damit bestehen die eingesetzten Testaufgaben jeweils aus einem Aufgabenstamm und einem oder mehreren Items. Die Items bestehen wiederum aus einem Itemstamm, die eine operationalisierte Fragestellung als Handlungsaufforderung enthalten.

Für die Operationalisierung und Aufgabenkonstruktion wurde im Rahmen der Pilotierung ein ausführlicher Aufgabenstamm mit Glossar zur Definition und Beschreibung der verwendeten Fachbegriffe entwickelt. Im Itemstamm wurde auf das Glossar der verwendeten Fachbegriffe verwiesen. Somit sind im Aufgaben- und Itemstamm alle Informationen bzgl. des biologischen Fachwissens enthalten, die für die Bearbeitung der Aufgabe erforderlich sind. Damit wurde der Einfluss der Variable Fachwissen reduziert bis vermieden. Im Itemstamm sind zusätzlich Informationen über die gezeigte Repräsentation, die lösungsrelevant sind, aber von den Studierenden selbst nicht erschlossen werden müssen. Dagegen werden Kenntnisse zum Umgang mit MER nicht vorgegeben. Dadurch, dass die zur Lösung der Aufgabe erforderlichen Fachinformationen im Aufgaben- oder Itemstamm enthalten sind, können die Testaufgaben als vorwissensunabhängig im Hinblick auf Fachwissen bezeichnet werden. Jeder Itemstamm enthält wiederum einen Text und ein Bild oder mehrere Bilder, wobei sich die Informationen in Text und Bild einander ergänzen. Damit wurden bewusst redundante Informationen ausgeschlossen und die Kompetenz der Integration angesprochen. Aufgrund der Gestaltung offener Aufgaben, hat sich die Lese- und Bearbeitungszeit im Rahmen der Pilotierung auf ungefähr 60 Minuten bemessen. Die Items sollten die Studierenden unterschiedlicher Fachrichtungen zum aktiven Integrieren anregen, d.h. es wurde von ihnen eine aktive Lese- und Übersetzungsleistung gefordert. Entsprechend mussten alle Items den Repräsentationswechsel erfüllen, indem redundante Informationen weggelassen und erforderliche Informationen zur Integration herangezogen werden konnten.

Im Folgenden werden exemplarisch drei Testaufgaben zu den drei im Modell postulierten Teilkomponenten MERI, MERII und MERIII und dem Repräsentationstyp Diagramm vorgestellt sowie zum Aufgabenkontext Stoffwechsel angehören.

Die Anforderung einer Aufgabe zur *Informationsentnahmenahen Integration (MERI)* besteht darin, Informationen aus einem gegebenen Text und Diagramm zu entnehmen und dabei zu integrieren. In ABBILDUNG 10 ist die Aufgabe Anp03 dargestellt, in der es um die Fotosyntheserate von Orchideen geht.

	Informationsent- nahmenahme Integration (MERI)	Stoffwechsel	Diagramm																														
Quelle	Abitur Leistungskurs 2011 G9 Bayern STARK Verlag, S. 2005-32																																
Aufgabentitel	Anpassung an besondere Standorte																																
Aufgaben- stamm und Impuls	Für die Fotosynthese benötigen Pflanzen Wasser und Kohlenstoffdi- oxid. Deshalb haben sie für die Anpassung an extreme Standorte verschiedene Überlebensstrategien entwickelt. Auf einer Tour durch den Botanischen Garten erfährt Lisa, dass in Florida die Orchidee <i>Catasetum</i> auf Baumkronen wächst. In Trockenzeiten wirft <i>Cata- setum</i> ihre Blätter ab. Auf einer Informationstafel ist das untenste- hende Diagramm abgebildet. Es zeigt die Kohlenstoffdioxid-Auf- nahme bzw. -Abgabe von <i>Catasetum</i> mit und ohne Blätter in Abhän- gigkeit von der Uhrzeit.																																
			<p>Das Diagramm zeigt die Kohlenstoffdioxid-Aufnahme und -Abgabe von <i>Catasetum</i> mit und ohne Blätter über den Tag. Die Y-Achse ist mit 'Kohlenstoffdioxid/ Stunde in g/h' beschriftet und reicht von -10 bis 25. Die X-Achse ist mit 'Uhrzeit in h' beschriftet und reicht von 0 bis 24. Die Kurve für 'Catasetum mit Blätter' (durchgezogene Linie) zeigt eine positive Aufnahme, die ab ca. 6 Uhr beginnt, ein Maximum von ca. 20 g/h bei 15 Uhr erreicht und abends wieder auf 0 sinkt. Die Kurve für 'Catasetum ohne Blätter' (gestrichelte Linie) zeigt eine negative Abgabe, die abends beginnt, ein Minimum von ca. -5 g/h bei 12 Uhr erreicht und abends wieder auf 0 sinkt.</p> <table border="1"> <caption>Datenpunkte aus dem Diagramm</caption> <thead> <tr> <th>Uhrzeit (h)</th> <th>Katasetum mit Blättern (g/h)</th> <th>Katasetum ohne Blätter (g/h)</th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td></tr> <tr><td>3</td><td>0</td><td>-2</td></tr> <tr><td>6</td><td>0</td><td>-1</td></tr> <tr><td>9</td><td>0</td><td>-1</td></tr> <tr><td>12</td><td>5</td><td>-5</td></tr> <tr><td>15</td><td>20</td><td>-2</td></tr> <tr><td>18</td><td>10</td><td>-1</td></tr> <tr><td>21</td><td>5</td><td>-1</td></tr> <tr><td>24</td><td>0</td><td>0</td></tr> </tbody> </table>	Uhrzeit (h)	Katasetum mit Blättern (g/h)	Katasetum ohne Blätter (g/h)	0	0	0	3	0	-2	6	0	-1	9	0	-1	12	5	-5	15	20	-2	18	10	-1	21	5	-1	24	0	0
Uhrzeit (h)	Katasetum mit Blättern (g/h)	Katasetum ohne Blätter (g/h)																															
0	0	0																															
3	0	-2																															
6	0	-1																															
9	0	-1																															
12	5	-5																															
15	20	-2																															
18	10	-1																															
21	5	-1																															
24	0	0																															
Fragestellung	Vergleichen Sie für Lisa die beiden Kurvenverläufe. Begründen Sie die Aufnahme und Abgabe von Kohlenstoffdioxid bei der Orchideenart <i>Catasetum</i> mit und ohne Blätter.																																
Antwort	<hr/>																																

ABBILDUNG 10: MODELLKONFORMES AUFGABENBEISPIEL ZUM UMGANG MIT MER, TEILKOMPETENZ INFORMATIONSENTNAHMENAHME INTEGRATION (MERI)

In einem ersten Schritt müssen die Informationen aus dem Diagramm entnommen werden, indem die beiden Kurvenverläufe für *Catasetum* mit und ohne Blätter beschrieben werden. Daraufhin folgt der Vergleich, dass bei Dunkelheit nur eine Kohlenstoffdioxidabgabe stattfindet, während bei optimaler Beleuchtung die Kohlenstoffdioxid-Aufnahme ab ca. 0,5°C ansteigt und bei ca. 8°C ein Maximum erreicht. Die Integrationsleistung besteht darin, dass bei

Kohlenstoffdioxid-Aufnahme vorrangig Fotosynthese stattfindet, während im Dunkeln keine Fotosynthese, sondern lediglich Zellatmung stattfindet.

Um zum Beispiel die Teilkompetenz *Konstruktionsnahe Integration (MERII)* im Aufgabenkontext Stoffwechsel mit einem Diagramm zu überprüfen, wurden Items konstruiert, in denen Informationen aus einem Text und einem Diagramm oder einem Text und einer Tabelle erschlossen werden müssen.

ABBILDUNG 11 zeigt die Aufgabe Epo05, in der Erythropoetin (EPO) als Dopingmittel zur Steigerung der körperlichen Leistungsfähigkeit thematisiert wird. Im Text wird erläutert, welche Eigenschaften das Dopingmittel Erythropoetin hat und wie der Mensch auf körperliche Leistungsanstrengung reagiert. Das abgebildete Diagramm zeigt den Anstieg der Konzentration der roten Blutkörperchen, wenn unterschiedliche EPO-Dosierungen verabreicht wurden. Die unabhängige Variable ist als Behandlungszeit in Wochen auf der X-Achse im Diagramm dargestellt. Im Beispielitem sollen die Studierenden die Hämatokrit-Werte nach einer Behandlungszeit von sechs Wochen ablesen und diese in ein neues Diagramm übertragen. Dabei soll der Hämatokrit-Wert in Abhängigkeit der EPO-Dosierung nach dieser Behandlungszeit dargestellt werden. Als richtige Antwort wird ein Liniendiagramm erstellt, da es sich um metrische Daten handelt.

	Konstruktions- nahe Integration (MERII)	Stoffwechsel	Diagramm																																																																																										
Quelle	Abitur Leistungskurs 2010 Bayern STARK Verlag, S.2010-2; Unterricht Biologie 309																																																																																												
Aufgabentitel	Die Wirkung von EPO																																																																																												
Aufgaben- stamm und Impuls	<p>Erythropoetin (EPO) ist ein typisches Dopingmittel zur Steigerung der körperlichen Leistungsfähigkeit. Es regt die Produktion der roten Blutkörperchen an. Als Maß für den Anteil der roten Blutkörperchen dient der Hämatokrit-Wert.</p> <p>Wird bei körperlicher Leistung sehr viel Sauerstoff verbraucht, beginnen die Nieren mit der Produktion von körpereigenem EPO. Seit 1969 kann EPO im Labor hergestellt werden und wird z.B. gegen Blutarmut bei Nierenkranken eingesetzt. Ziel ist es, den Mangel an dem Hormon Erythropoetin auszugleichen und so die Blutarmut zu beheben.</p> <p>Lisa entdeckt in ihrem Schulbuch das nebenstehende Diagramm. Es zeigt den Anstieg der Konzentration der roten Blutkörperchen nach Verabreichung unterschiedlicher EPO-Dosierungen in Abhängigkeit der Behandlungszeit in Wochen.</p>																																																																																												
			<table border="1"> <caption>Data from the graph in the text</caption> <thead> <tr> <th>Behandlungszeit in Wochen</th> <th>15 U/kg</th> <th>50 U/kg</th> <th>150 U/kg</th> <th>500 U/kg</th> <th>750 U/kg</th> </tr> </thead> <tbody> <tr><td>0</td><td>18</td><td>18</td><td>18</td><td>18</td><td>18</td></tr> <tr><td>1</td><td>19</td><td>20</td><td>22</td><td>25</td><td>28</td></tr> <tr><td>2</td><td>20</td><td>22</td><td>25</td><td>30</td><td>34</td></tr> <tr><td>3</td><td>21</td><td>24</td><td>28</td><td>35</td><td>40</td></tr> <tr><td>4</td><td>22</td><td>26</td><td>31</td><td>40</td><td>46</td></tr> <tr><td>5</td><td>23</td><td>28</td><td>34</td><td>45</td><td>52</td></tr> <tr><td>6</td><td>24</td><td>30</td><td>37</td><td>50</td><td>58</td></tr> <tr><td>7</td><td>25</td><td>32</td><td>40</td><td>55</td><td>64</td></tr> <tr><td>8</td><td>26</td><td>34</td><td>43</td><td>60</td><td>70</td></tr> <tr><td>9</td><td>27</td><td>36</td><td>46</td><td>65</td><td>76</td></tr> <tr><td>10</td><td>28</td><td>38</td><td>49</td><td>70</td><td>82</td></tr> <tr><td>11</td><td>29</td><td>40</td><td>52</td><td>75</td><td>88</td></tr> <tr><td>12</td><td>30</td><td>42</td><td>55</td><td>80</td><td>94</td></tr> <tr><td>13</td><td>31</td><td>44</td><td>58</td><td>85</td><td>100</td></tr> </tbody> </table>	Behandlungszeit in Wochen	15 U/kg	50 U/kg	150 U/kg	500 U/kg	750 U/kg	0	18	18	18	18	18	1	19	20	22	25	28	2	20	22	25	30	34	3	21	24	28	35	40	4	22	26	31	40	46	5	23	28	34	45	52	6	24	30	37	50	58	7	25	32	40	55	64	8	26	34	43	60	70	9	27	36	46	65	76	10	28	38	49	70	82	11	29	40	52	75	88	12	30	42	55	80	94	13	31	44	58	85	100
Behandlungszeit in Wochen	15 U/kg	50 U/kg	150 U/kg	500 U/kg	750 U/kg																																																																																								
0	18	18	18	18	18																																																																																								
1	19	20	22	25	28																																																																																								
2	20	22	25	30	34																																																																																								
3	21	24	28	35	40																																																																																								
4	22	26	31	40	46																																																																																								
5	23	28	34	45	52																																																																																								
6	24	30	37	50	58																																																																																								
7	25	32	40	55	64																																																																																								
8	26	34	43	60	70																																																																																								
9	27	36	46	65	76																																																																																								
10	28	38	49	70	82																																																																																								
11	29	40	52	75	88																																																																																								
12	30	42	55	80	94																																																																																								
13	31	44	58	85	100																																																																																								
Fragestellung	Konstruieren Sie ein vollständig beschriftetes Diagramm, das den Hämatokrit-Wert in Abhängigkeit von der EPO-Dosierung nach einer Behandlungsdauer von sechs Wochen zeigt.																																																																																												
Antwort																																																																																													

ABBILDUNG 11: MODELLKONFORMES AUFGABENBEISPIEL ZUM UMGANG MIT MER, TEILKOMPETENZ KONSTRUKTIONSDNAHE INTEGRATION (MERII)

Im Rahmen einer *Translation und Transformation aus mehreren Repräsentationen (MERIII)* ist die Anforderung verknüpft, dass aus einem gegebenen Text und beispielsweise einem Diagramm, ein neues Diagramm konstruiert und damit transformiert werden muss. Als Beispiel wird im Folgenden das Item See09 ausgeführt (vgl. ABBILDUNG 12).

	Translation und-Transformation (MERIII)	Stoffwechsel	Diagramm
Quelle	NATURA Oberstufe Gesamtband Klett Verlag, S. 172		
Aufgabentitel	Das Ökosystem See		
Aufgabenstamm und Impuls	<p>Durch menschliche Einflüsse wie z.B. Dünger können vermehrt Nährstoffe in Gewässer gelangen. Diese führen zu einer erhöhten Produktion von Algen. Beim Abbau der Biomasse durch Mikroorganismen wird Sauerstoff verbraucht und der Sauerstoffgehalt des Sees verändert sich. In größeren Tiefen können Pflanzen wegen des Lichtmangels keine Fotosynthese mehr betreiben. Es tritt Sauerstoffmangel ein, die Biomasse nimmt zu und am Boden bildet sich Faulschlamm.</p> <p>Während eines Forschungspraktikums hat Paul die untenstehenden Diagramme zum Waschsee erstellt. Paul fragt sich, was während der einzelnen Monate mit dem See geschehen ist. Er hat in seinem Praktikum gelernt, dass eine Algendichte von 5000 pro ml der Faulschlammdicke von 1m entspricht.</p>		
Fragestellung	Berechnen Sie für Paul die Dicke des Faulschlammes an der Wasseroberfläche für die Monate März, Mai und August mithilfe der Diagramme. Konstruieren Sie anschließend mit den erhaltenen Werten ein vollständig beschriftetes Diagramm.		
Antwort			

ABBILDUNG 12: MODELLKONFORMES AUFGABENBEISPIEL ZUM UMGANG MIT MER, TEILKOMPETENZ TRANSLATION UND TRANSFORMATION (MERIII)

Im Rahmen dieser MERIII-Integration ist es erforderlich, dass zunächst die Informationen aus dem Text und den drei Diagrammen entnommen werden muss. In diesem Zusammenhang ist es erforderlich die Algendichte korrekt abzulesen und mit der Information aus dem Text bezüglich der Faulschlammdicke zu integrieren (März: 0,2 m, Mai, 2 m, August: 4 m). Anschließend muss die Wahl des korrekten Diagrammtyps, ein Säulendiagramm für die drei Monate, getroffen und die richtigen Werte eingetragen werden.

Antwortformate: Bei der Itementwicklung ist die Wahl des Antwortformats entscheidend. Es beschreibt die Art und Weise wie eine Aufgabe beantwortet wird (vgl. Lienert & Raatz, 1998). Dabei kann zwischen offenen, halboffenen und geschlossenen Aufgaben unterschieden werden (Graf, 2001). Geschlossene oder auch gebundene Aufgaben zeichnen sich dadurch aus, dass die Testperson aus einer gegebenen Antwortauswahl eine oder mehrere auswählen muss. Dieses Antwortformat ist auch unter dem Stichwort Multiple-Choice-Aufgabe bekannt (Fischer & Draxler, 2001). In der psychologischen Diagnostik werden vor allem geschlossene Antwortformate eingesetzt, da sie zeitlich ökonomisch sind und von einer hohen Auswertungsobjektivität profitieren. In der vorliegenden Studie werden aus diesem Grund Multiple-Choice-Aufgaben ebenfalls eingesetzt. Dabei wurde darauf geachtet, dass die Testitems im Multiple-Choice-Format „2 aus 5“ konstruiert wurden. Die Studien von Kubinger et al. (2010) und Hohensinn & Kubinger (2011) konnten zeigen, dass sich solche Testitems als nicht einfacher erwiesen wie jene im offenen Antwortformat und sich der offenen Antwortkonstruktion ähneln. Zum gleichen Ergebnis kommen Kubinger & Gottschall (2007), indem das Antwortformat „x aus 5“ ebenfalls keinen signifikanten Schwierigkeitsunterschied zeigte (vgl. Vock, 2009). Des Weiteren sind Testaufgaben im offenen Antwortformat enthalten. Das offene oder auch freie Antwortformat beschreibt das Antworten ohne Antwortmöglichkeiten (Lienert & Raatz, 1998). Im Rahmen der vorliegenden Untersuchung kommen neben Kurzantworten auch Aufgaben mit erweitertem Antwortformat vor (vgl. Klieme et al., 2000; Kulgemeyer, 2009). In diesen Aufgaben wird die Testperson dazu aufgefordert, eine Antwort oder Lösung frei zu formulieren und schriftlich zu dokumentieren (Hartig & Jude, 2007). Testaufgaben unterscheiden sich in Bezug auf die Offenheit des erwarteten Lösungsverhaltens dadurch, dass die Lösungswege unterschiedlich „offen“ variiert werden (Martinez, 1999). Das verwendete offene Aufgabenformat orientiert sich dabei an charakteristischen Prüfungssituationen im schulischen Bereich und der schriftlichen Abiturprüfung (Jankisz & Moosbrugger, 2008, zitiert nach Florian, 2012). Entscheidend hierzu sind Untersuchungen, die zeigen, dass offene und geschlossene Antwortformate unterschiedliche kognitive Prozesse bei der Beantwortung verlangen, aber auch unterschiedlich die Aufgabenschwierigkeit beeinflussen (Klieme et al.,

2010; Martinez, 1999). Dadurch wird der Prozess der Problemlösung während der Aufgabebearbeitung transparent (Duit, Häußler & Prenzel, 2002). Denk- und Argumentationsweisen können beleuchtet werden sowie differenzierte Beschreibungen hinsichtlich qualitativ unterschiedlicher Verständnisstufen (Hammann & Jördens, 2014). Dabei hängt die Qualität oder Korrektheit der Lösung von den Beurteilern ab, wofür eine sorgfältige Dokumentation der Beurteilungskriterien essentiell ist (Hartig & Jude, 2007). Studien zeigen, dass offene Testitems mit freien, langen Antworten tendenziell schwieriger zu lösen sind als Multiple-Choice-Items (vgl. Lachmayer, 2008; Prenzel et al., 2002; Zimbardo & Gerrig, 1999). Während bei offenen Aufgaben ein „aktives Hervorbringen und Konstruieren von Bedeutungszusammenhängen erforder[lich]“ ist (Hammann & Jördens, 2014, S. 170), kann bei Multiple-Choice-Aufgaben durch systematisches Ausschließen von falschen Antworten die richtige Lösung ermittelt werden, dadurch ist eine gewisse Verrechnungssicherheit gegeben (Vock, 2009; Bühner, 2011). Die Vielgestaltigkeit der Antworten führt hingegen zu einer erschwerten Auswertung, da offene Aufgaben zeitaufwändiger zu codieren sind, aber dennoch die Gütekriterien Objektivität und Reliabilität erfüllen sollen (Hammann & Jördens, 2014). Beim offenen Antwortformat hängt das richtige Lösungsverhalten zu einem gewissen Teil vom Testleiter ab, der das Antwortverhalten von Testpersonen in Leistungstests unterschiedlich richtig oder falsch bewerten kann. Vock (2009) weist darauf hin, dass Kategoriensysteme niemals erschöpfend sein können, wodurch ein Handlungsspielraum bei der Bewertung der Antworten resultiert und die Reliabilität der Messung eingeschränkt werden kann.

Diesen Untersuchungen stehen hingegen stets fachliche und inhaltliche Überlegungen gegenüber. Da in der vorliegenden Studie zum Umgang mit MER vor allem Wissensanforderungen verlangt werden, die eine freie und selbstständige Interpretation bzw. Konstruktion erfordern, wird das offene Antwortformat bevorzugt eingesetzt. Des Weiteren ist das offene Antwortformat charakteristisch für Prüfungen im schulischen Bereich (Jankisz & Moosbrugger, 2007) sowie in der Abiturprüfung (Kühn, 2010; Florian, Sandmann, & Schmiemann, 2014) und entspricht damit dem gängigen Lösungsverhalten von Schülerinnen und Schülern sowie Studierenden im ersten Semester. Eine sorgfältige Aufgabenkonstruktion, Dokumentation und Beurteilung über ein Kategoriensystem sichern die psychologischen Kriterien einer validen Messung. Aus diesem Grund wurden die Anforderungen an die Aufgabenlösung konkretisiert und in einem Erwartungshorizont festgehalten, was einer Strukturierung der Antwort sowie Spezifizierung der Aufgabenstellung diene. Die Nachteile offener Aufgaben, insbesondere der hohe zeitliche Aufwand für die Beantwortung sowie eine aufwendigere Auswertung mit unter Umständen eingeschränkter Objektivität werden für die Zielsetzung der vorliegenden Untersuchung in Kauf genommen.

Den Testpersonen wurden zur Gewährleistung einer hohen Auswertungsökonomie und –Auswertungsobjektivität (vgl. Bühner, 2006) Testitems in drei Antwortformaten (offen – Kurzantwort, offen – ausführliche Antwort, geschlossen – Multiple-Choice) vorgelegt. Die Verteilung der Items auf die unterschiedlichen Antwortformate in der Vor- und Hauptstudie sind in TABELLE 2 dargestellt. In der Vorstudie besaßen etwa 60% der Testitems ein offenes ausführliches Antwortformat, das andere gute Drittel wurde annähernd gleich auf Testitems mit offenem Kurzantwortformat und geschlossenem Multiple-Choice-Format aufgeteilt. Pro Testitem liegen zwischen 15 und 62 Antworten vor.

TABELLE 2: VORSTUDIE - ITEMS IN UNTERSCHIEDLICHEN ANTWORTFORMATEN

	Vorstudie							
	MERI		MERII		MERIII		Gesamt	
	N	%	N	%	N	%	N	%
geschlossen – Multiple Choice	25	11,90	/	0	10	4,76	35	16,66
offen – Kurzantwort (Kurzantwort-, Ablese-, Zuordnungsaufgaben)	16	7,62	24	11,43	8	3,81	48	22,86
offen – ausführliche Antwort (Operatoren)	54	25,71	41	19,52	32	15,24	127	60,48
Gesamt	95	45,24	65	30,95	50	23,81	210	100

Vorwissen: Die Testaufgaben der vorliegenden Studien wurden so konstruiert, dass kein inhaltliches Vorwissen zur Lösung der Aufgabe benötigt wurde. Damit konnten auch Testpersonen ausgewählt werden, die Biologie nicht als Studienfach haben, was eine breite Auswertung sicherte. Die Aufgaben wurden mit dem erforderlichen Fachwissen konstruiert, sodass alle Testpersonen die gleichen Voraussetzungen vorzufinden hatten. Damit wurden standardisierte Bedingungen geschaffen, um ausschließlich das Wissen beim Umgang mit unterschiedlichen MER zu erfassen.

Itemumfang: Insgesamt wurden 210 Items entwickelt und im Rahmen der Vorstudie eingesetzt. Unter Beachtung aller sinnvollen Kombinationsmöglichkeiten umfasst das Kompetenzstrukturmodell insgesamt 18 Zellen einer dreidimensionalen Matrix. Diese erschließt sich aus den drei Integrationsleistungen der drei Kompetenzaspekte *MERI*, *MERII* und *MERIII*, den

drei *Aufgabenkontexten* Ökologie, Stoffwechsel, und Genetik sowie den zwei *Repräsentationstypen* Diagramm und Schemazeichnung. Aufgrund der postulierten Annahme des Kompetenzmodells ist diese insgesamt sehr hohe Anzahl an Testaufgaben notwendig, wenn die Zellen der Matrix für alle Kombinationen ausreichend besetzt werden sollen und von einer entsprechenden Itemselektion und damit Itemreduktion auszugehen ist. Zur Operationalisierung des Kompetenzmodells müssen also 18 Zellen für eine trennscharfe Abbildung der einzelnen Dimensionen mit Testaufgaben gefüllt werden (vgl. TABELLE 3).

TABELLE 3: ITEMVERTEILUNG VORSTUDIE (N=210)

Umgang mit multiplen externen Repräsentationen																	
MERI						MERII						MERIII					
Öko		Stw		Gen		Öko		Stw		Gen		Öko		Stw		Gen	
41		23		23		31		23		18		18		16		16	
Items		Items		Items		Items		Items		Items		Items		Items		Items	
D	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S
29	12	16	7	10	13	20	11	14	9	10	8	11	7	10	6	8	8
95 Testaufgaben						65 Testaufgaben						50 Testaufgaben					

4.3 TESTHEFTDESIGN

Zur Untersuchung des Umgangs mit multiplen externen Repräsentationen in unterschiedlichen biologischen Aufgabenkontexten und anhand unterschiedlicher Repräsentationstypen wurde, in Anlehnung an Large-Scale-Assessments, ein Studiendesign mit unterschiedlichen Testheften realisiert. Damit die Testzeit für die Studierenden möglichst geringgehalten werden kann, weist die Zusammenstellung der einzelnen Testhefte einige Besonderheiten auf. Ein vollständiges Facettendesign, welches alle Komponenten für alle Dimensionen kombiniert, wurde aufgrund der großen Anzahl benötigter Aufgaben nicht systematisch durchgeführt (vgl. Schmiemann, 2010; Einhaus, 2007). Aufgrund der gezielten Variation von Aufgabenmerkmalen bei Kontrolle anderer, kann dennoch von einer empirisch zufriedenstellenden systematischen Erfassung der Merkmale ausgegangen werden. Da die Vorstudie primär die Überprüfung der modellkonform konstruierten Items auf ihre Eignung als Testaufgaben verfolgt, ist dies aus psychometrischer Hinsicht hinnehmbar.

Für die Verteilung der Items auf die verschiedenen Testhefte wurde im Rahmen der Vorstudie ein Youden-Square-Design (vgl. Frey & Annageldyev, 2013; 2015) verwendet. Da bei Kom-

petenztestungen auf die individuellen Ausprägungen eines interessierenden latenten Konstrukts geschlossen werden soll und für eine valide Abbildung von Merkmalen eine hohe Anzahl an Testitems vorgelegt werden muss, als man in einer vertretbaren Testzeit vorgeben kann, sind Testhefte eine effiziente Alternative zur vollständigen Permutation (Frey, Hartig, & Rupp, 2009). Ähnlich dem Multi-Matrix-Design, das z.B. in PISA oder bei NAEP Anwendung findet, müssen die Probanden nicht alle Testaufgaben bearbeiten, d.h. es wird ihnen nur eine begrenzte Anzahl und Auswahl an Testaufgaben aus dem vollständigen Itempool vorgelegt. Der Datensatz weist damit designbedingte Datenlücken auf (Rost, 2004) und erfordert eine Auswertung mit Methoden der probabilistischen Testtheorie, bei der die fehlenden Messwerte geschätzt werden (s. Kapitel 5.2). Daraus folgt, dass insgesamt 31 Testhefte mit 31 Clustern konstruiert wurden, wobei in jedem Testheft sechs Cluster repräsentiert werden und jeder Cluster an allen sechs Positionen des Testhefts einmal vorkommt. Die Anzahl der Items pro Testheft wurde so gewählt, dass alle Probanden die Testhefte innerhalb einer Vorlesungsstunde zwischen 60-90 Minuten beantworten können. Hierfür wurden alle Testitems zu verschiedenen Clustern, d.h. einer Gruppe von Items, zusammengefasst, die wiederum in die einzelnen Testhefte fließen (gemäß Youden-Square, vgl. Frey, Hartig, & Rupp, 2009). Youden-Square-Designs (YSD) sind spezielle balancierte unvollständige Blockdesigns (BIBDs), die wiederum ein binäres, unvollständiges Blockdesign für eine bestimmte Anzahl an Clustern in einer bestimmten Anzahl an Testheften abbilden (Frey, Hartig, & Rupp, 2009). Für BIBDs gilt, dass die Anzahl der Cluster pro Testheft kleiner ist, als die Anzahl aller Cluster. Des Weiteren gelten folgende Voraussetzungen:

- Jedes Cluster (t) kommt höchstens einmal in einem Testheft (b) vor.
- Jedes Cluster erscheint gleich häufig (r) über alle Testhefte.
- Jedes Testheft hat eine identische Länge und enthält die gleiche Anzahl an Clustern (k).
- Jedes Paar von Clustern tritt zusammen in Testheften mit der gleichen Frequenz (λ) auf.

Die Parameter t , b , r , k und λ sind die Parameter des Designs und charakterisieren das unvollständige Blockdesign (vgl. Frey, Hartig, & Rupp, 2009). Youden-Square-Designs unterscheiden sich von BIBDs durch die zusätzliche Bedingung, dass jedes Cluster auf jeder Position des Testhefts mit gleicher Frequenz vorkommen muss. Daraus ergibt sich, dass die Anzahl der Testhefte gleich der Anzahl der Cluster ($t=b$) ist. Zusätzlich ist die Häufigkeit, mit der ein Cluster über alle Testhefte verteilt vorkommt, gleich der Anzahl der Positionen pro Testheft ($k=r$). Dadurch lassen sich Positionseffekte kontrollieren, gleichzeitig existieren YSD nur für eine spezielle Kombination der Designparameter.

Im Rahmen der Vorstudie wurde mithilfe des Java-Programms *Youden* (Frey & Annageldyev, 2015) das Design der vorliegenden Arbeit durch die Anzahl der Testhefte und die Anzahl der Cluster pro Testheft sowie die Verteilung der Cluster festgelegt. Die gewählten Designparameter waren $t=b=31$ und $r=k=6$ sowie $\lambda=1$. In diesem Zusammenhang wurden jeweils 18 Items aus dem Itempool ausgewählt und auf 62 Testhefte einer A- und B-Serie verteilt. Das Testheft-Design der Vorstudie ist in TABELLE 4 dargestellt.

TABELLE 4: VORSTUDIE - TESTHEFTDESIGN ANGELEGT ALS YOUDEN-SQUARE-DESIGN (210 ITEMS)

TH 1	TH 2	TH 3	TH 4	TH 5	TH 6	TH 7	TH 8	TH 9	TH 10	TH 11
2	3	4	5	6	7	8	9	10	11	12
6	7	8	9	10	11	12	13	14	15	16
12	13	14	15	16	17	18	19	20	21	22
25	26	27	28	29	30	31	1	2	3	4
26	27	28	29	30	31	1	2	3	4	5
28	29	30	31	1	2	3	4	5	6	7
TH 12	TH 13	TH 14	TH 15	TH 16	TH 17	TH 18	TH 19	TH 20	TH 21	TH 22
13	14	15	16	17	18	19	20	21	22	23
17	18	19	20	21	22	23	24	25	26	27
23	24	25	26	27	28	29	30	31	1	2
5	6	7	8	9	10	11	12	13	14	15
6	7	8	9	10	11	12	13	14	15	16
8	9	10	11	12	13	14	15	16	17	18
TH 23	TH 24	TH 25	TH 26	TH 27	TH 28	TH 29	TH 30	TH 31		
24	25	26	27	28	29	30	31	1		
28	29	30	31	1	2	3	4	5		
3	4	5	6	7	8	9	10	11		
16	17	18	19	20	21	22	23	24		
17	18	19	20	21	22	23	24	25		
19	20	21	22	23	24	25	26	27		

Die aus dem Youden-Square-Design resultierende Datenmatrix zeigt eine große Anzahl an fehlenden Werten auf. Die Aufgabenblöcke in diesem Design sind jedoch derart verlinkt, dass mit Methoden der probabilistischen Testtheorie dennoch eine zuverlässige Schätzung der Personenparameter ermöglicht wird (s. Kapitel 5.2.2). Obwohl es zu einer unterschiedlichen Bearbeitung der Testaufgaben kommt, können die erfassten Kompetenzen für alle getesteten Studierenden auf einer gemeinsamen Skala abgebildet werden (Rauch & Hartig, 2008).

In einem nächsten Schritt folgt die Zuordnung der Items zu den Clustern. Die Verteilung der 210 konstruierten Testaufgaben auf die berechneten 31 Cluster würde zu Testheften führen, die ca. 40 Testitems enthalten. Die Beantwortung solcher Testhefte hätte zu einer langen Testzeit geführt und gleichzeitig die Gefahr einer kognitiven Überlastung der teilnehmenden Studierenden bedingt. Aus diesem Grund wurden für die 31 Testhefte zwei Serien bestimmt, eine Testheftserie A (bestehend aus 31 Testheften) sowie eine Testheftserie B (bestehend aus 31 Testheften). Die Verteilung der Testaufgaben erfolgt so, dass ausgewählte Items über die beiden Serien verlinkt sind. Beispielsweise wurde in der A-Serie das Item zur *Informationsentnahmenahen Integration (MERI)* präsentiert, während der gleiche Aufgabenstamm in der B-Serie verwendet wurde und die Aufgabe entsprechend als *Konstruktionsnahe Integration (MERII)* formuliert wurde. Damit konnte gleichzeitig eine Konfundierung von Testaufgaben vermieden werden, da im Rahmen der Aufgabenkonstruktion häufig für die Komponenten MERI und MERII gleiche Diagramme verwendet wurden, einmal als informationsentnahmenahen und einmal als konstruktionsnahe Integration formuliert. So wurden zum Beispiel für den Aufgabenstamm *Anpassung an besondere Standorte* diejenigen Items getrennt, die das gleiche Diagramm für unterschiedliche Fragestellungen verwendeten. Da diese Testaufgabe dieselbe Information enthält, durften solche Aufgaben nicht zusammen in ein Cluster und nicht in ein Testheft, sondern waren Ausgangspunkt für die Itemverteilung in die Testheftserie A und B. Daneben wurde darauf geachtet, dass die einzelnen Cluster der jeweiligen A- und B-Serie vergleichbar sind, d.h. beispielsweise den gleichen Inhalt haben und sich in der MER-Integration über mindestens zwei Teilkomponenten unterscheiden. Auf dieser Basis musste die Itemanzahl auf 186 Testaufgaben beschränkt werden, damit in jedem der insgesamt 62 Testhefte genau 3 Items pro Cluster eingruppiert werden konnten und eine Verlinkung der Cluster ermöglicht wird (vgl. Anhang 14.1). Damit gehen 186 Testaufgaben in den Itempool der vorliegenden Untersuchung ein. Durchschnittlich wurde jede Aufgabe von 15 bis 62 Personen bearbeitet und beantwortet. In Serie A lag die durchschnittliche Beantwortung pro Item bei $n=42,7$ Antworten, in Serie B bei $n=36,94$ Antworten.

4.4 DURCHFÜHRUNG DER VORSTUDIE

Die Pilotierung der Aufgaben wurde als Paper-and-Pencil-Test durchgeführt, da es ein etabliertes und ökonomisches Verfahren zur Kompetenzmessung ist (Stecher & Klein, 1997; Hartig & Jude, 2007). Die Durchführung fand statt im Zeitraum April – Juli und Oktober 2014. Darüber hinaus wurden für die empirische Beschreibung von Repräsentationskompetenz Informationen über die Bearbeitungszeit gesammelt, die für die Beantwortung der Items benötigt wurde (vgl. Osterlind, 1998).

Die Durchführung der Vorstudie fand im Rahmen einer Vorlesungs- oder Übungsstunde innerhalb der Universität statt. Wenn möglich wurde dafür ausreichend Bearbeitungszeit zur Verfügung gestellt oder die Untersuchung fand in den letzten 30-45 Minuten der Zeit statt. Dabei wurde darauf geachtet, dass die Studierenden im Anschluss keine weitere Vorlesung, Übung oder ein Praktikum hatten, um ausreichend zusätzliche Zeit zur Verfügung zu stellen. Zum Ausfüllen der Fragebögen hatten die Studierenden in der Pilotierung bis zu 90 Minuten Zeit (davon 60 Minuten für die Bearbeitung der Testaufgaben). Für eine genauere Einschätzung der benötigten Zeit wurde diese zusätzlich durch die Testleitung gemessen und notiert. Anlass und Zweck der Untersuchung wurden zu Beginn durch den vortragenden Dozenten oder durch die Testleitung selbst kurz erläutert.

5 STATISTISCHE METHODEN DER VORSTUDIE

In diesem Kapitel werden die Hintergründe für die statistische Analyse zum entwickelten Kompetenzstrukturmodell erläutert. Für die Analyse wurden Verfahren aus der Klassischen Testtheorie (KTT) und Probabilistischen Testtheorie (PTT) verwendet. Die Analyseverfahren ergänzen sich dadurch, dass die KTT Messwerte als Voraussetzung während die PTT Messwerte erst als Ergebnis einer Testanalyse versteht (Rost, 2004).

Die Schätzung des Repräsentationswissens und die Modellierung des mehrdimensionalen Kompetenzstrukturmodells zum Umgang mit MER werden mit Modellen der PTT vorgenommen, da sie hierzu geeignete Analysewerkzeuge bereitstellen (Hartig & Frey, 2013). Weitere Analysen in Bezug auf den Zusammenhang von Personenmerkmalen und Repräsentationswissen werden mit Verfahren der KTT durchgeführt. Die Analysen aus der probabilistischen Testtheorie werden mit dem Programm *Conquest 2.0* (Wu, Adams & Wilson, 2007) durchgeführt. Die klassischen Analysen im Rahmen der KTT werden mit *IBM SPSS 22-24* (IBM SPSS Statistics 22.0 bis 24.0) ausgeführt.

5.1 KLASSISCHE UND PROBABILISTISCHE TESTTHEORIE

Testtheorien befassen sich mit der Frage, inwieweit ein Zusammenhang zwischen Testverhalten und dem zu erfassenden psychischen Merkmal besteht (Bühner, 2006; Rost, 2004). Die klassische und probabilistische Testtheorie stellen zwei unterschiedliche Ansätze zur Datenanalyse dar und bieten Verfahren zur Schätzung von Itemschwierigkeiten, Personenfähigkeiten sowie zur Testgüte (Magno, 2009). Den beiden Testtheorien ist hingegen gemein, dass die beobachtbaren Messwerte mit einer Personeneigenschaft in Beziehung gesetzt werden. Im Kontext psychometrischer Modelle spricht man von so genannten latenten Variablen, die nicht unmittelbar beobachtet werden können, sondern über das Testverhalten erschlossen werden müssen (Hartig & Klieme, 2006). Durch das Zusammenfassen der Antworten auf eine Menge von Aufgaben zu einem einzelnen Testwert, können Personen als unterschiedlich fähig beschrieben werden (Hartig & Jude, 2007). Sowohl die klassische als auch die probabilistische Testtheorie erlauben also Rückschlüsse vom Antwortverhalten einer Versuchsperson im Test auf ein bestimmtes Merkmal. Laut klassischer Testtheorie (KTT) werden die erhobenen Testwerte unmittelbar auf messfehlerbehaftete Merkmalsausprägungen von Personen zurückgeführt, d.h. dass sich der beobachtete Messwert einer Person durch einen „wahren“ Wert für ihre Fähigkeit und einem Messfehler zusammensetzt (u.a. Embretson & Reise, 2000; Wilson, 2005; Bortz & Döring, 2006). Die richtigen Antworten bzw. die Zahl der richtigen

Antworten, gemessen als Summenwert, werden als direktes Maß für die Personenfähigkeit angesehen (u.a. Embretson & Reise, 2000). Eine Schwäche der KTT im Rahmen von Kompetenzmessungen ist jedoch, dass keine expliziten Annahmen zwischen Personenmerkmal und Itembeantwortung formuliert werden.

Die probabilistische Testtheorie (PTT) geht davon aus, dass nicht beobachtbare Personeneigenschaften einen Einfluss auf das Testverhalten einer Person haben (vgl. ABBILDUNG 13). Die PTT spricht in diesem Zusammenhang nicht von direkt messbaren Persönlichkeitsmerkmalen, sondern von *latenten* Merkmalen, die nur indirekt über *manifeste* Variablen (z.B. Antworten in einem Test) erschlossen werden können. Damit kann mit Modellen der PTT auf ein dahinterliegendes latentes Konstrukt geschlossen werden, wobei die Personenparameter als Angabe über die Ausprägung des latenten Konstrukts interpretiert werden (Wu & Adams, 2007). Die beobachteten Antworten einer Person resultieren aus einer latenten Fähigkeit, die beobachtbare Zusammenhänge *produziert* (Bühner, 2006). Man schließt also vom Antwortverhalten auf das Personenmerkmal ohne dass man von direkt messbaren Persönlichkeitswerten ausgeht (vgl. Rost, 2004). Die Fähigkeit der Testperson bestimmt in Wechselwirkung mit der Aufgabenschwierigkeit wie wahrscheinlich es ist, eine Aufgabe korrekt zu lösen (Kauertz, 2014).

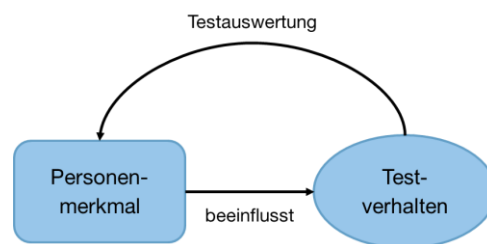


ABBILDUNG 13: ZUSAMMENHANG VON PERSONENMERKMAL UND TESTVERHALTEN (IN ANLEHNUNG AN ROST, 2004, S. 21)

Vor diesem Hintergrund kann eine Kompetenz, ausgedrückt als latente Variable θ , durch manifeste Variablen (Antworten auf die Testitems) beobachtet werden. Dieser Zusammenhang von latenter Variable und manifesten Variablen wird mittels verschiedener psychometrischer Testmodelle modelliert (vgl. Rost, 2004). ABBILDUNG 14 verdeutlicht diese Annahme.

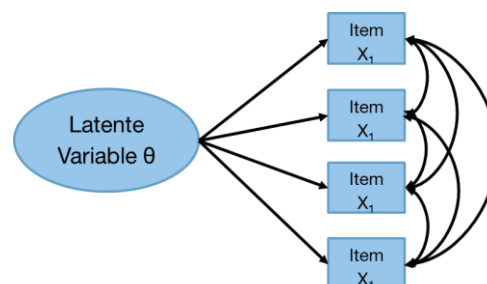


ABBILDUNG 14: LATENTE VARIABLE UND MANIFESTE MERKMALSAUSPRÄGUNGEN

Mathematische Modelle stellen einen Zusammenhang zwischen latenter Personeneigenschaft und dem manifesten, also beobachtbarem, Testverhalten her (u.a. Rost, 2004). Das Lösungsverhalten bei der Itembearbeitung steht also im Zusammenhang zum zugrundeliegenden Kompetenzkonstrukt (Klieme, Hartig, & Rauch, 2008). Die Messung der interessierenden Merkmale, d.h. der latenten Variablen, basiert auf einer Schätzung, die auf Basis der Zusammenhänge zwischen beobachtbaren Variablen und den Annahmen im Messmodell berechnet wird (Hartig & Klieme, 2006). Damit werden intervallskalierte, lineare Schätzer für die Personeneigenschaften gewonnen, wodurch statistische Analysen mit parametrischen Testverfahren ermöglicht werden (u.a. Boone & Scantlebury, 2006; Embretson & Reise, 2000).

Zusammenfassend betrachtet, zeigt die Rasch-Skalierung, wie gut die einzelnen Aufgaben ein gemeinsames Konstrukt, d.h. die latente Fähigkeit widerspiegeln. Dieses Konstrukt wird wiederum durch die Testaufgaben bzw. die Anforderung in dieser Aufgabe bestimmt. Die Operationalisierung des theoretischen Konstrukts erfolgt damit in Form von Aufgabenmerkmalen. Dadurch können vorab (*a priori*) theoretische Annahmen über die Beziehung zwischen Aufgabenmerkmalen und Personenfähigkeit aufgestellt werden, was die Validität des entwickelten Tests sichert (Kauertz, 2014).

5.2 GRUNDLAGEN DER PROBABILISTISCHEN TESTTHEORIE

Zunächst wird das eindimensionale Rasch-Modell als Analysemodell der Item Response Theory (IRT) vorgestellt. Auf die ausführliche Beschreibung von Modellgleichungen der verwendeten Rasch-Modelle wird im Folgenden verzichtet, da diese eine komplexe mathematische Herleitung erfordern. An dieser Stelle sei auf einschlägige Literatur verwiesen (z.B. Rost, 1996; Rost, 2004; Adams, Wilson & Wang, 1997).

Mit Modellen der IRT kann die Passung von angenommenen Modellstrukturen an empirische Daten und die systematische Varianz des zu messenden Merkmals getestet werden (Hartig & Frey, 2013). Das codierte Antwortverhalten auf die entwickelten Testaufgaben steht in Beziehung zum theoretischen Konstrukt (Wilson, 2005). Die Herausforderung bei der Entwicklung von Kompetenzmodellen ist die Modellierung von situativen Anforderungen und Personenmerkmalen, um die individuellen Testwerte als Fähigkeit zur Bewältigung spezifischer Anforderungen im Sinne des Kompetenzbegriffs interpretieren zu können (Hartig, 2008). Hierfür bieten sich Modelle der Item Response Theory (IRT) an (vgl. Rost, 2004), da sie durch den Vergleich von Personenparametern mit Itemparametern, abgebildet auf einer gemeinsamen Skala, eine kriteriumsorientierte Testwertinterpretation erlauben (vgl. Embretson & Reise, 2000) und eine Grundlage für die Beschreibung von Kompetenzniveaus ermöglichen (vgl.

Hartig & Klieme, 2006). Die Grenzen der Modellierung von Kompetenzen mit IRT-Modellen liegen bei einer Verletzung der lokalen stochastischen Unabhängigkeit, d.h. die Art der Beantwortung einer Testaufgabe ist unabhängig vom Lösungsverhalten einer anderen Testaufgabe im selben Test (Hartig & Frey, 2013) und darf nur vom zu erfassenden Konstrukt abhängen (Neumann, 2014). Dies bedeutet wiederum, dass das Rasch-Modell erst bei lokaler stochastischer Unabhängigkeit der Aufgaben gültig ist. Eine Schwäche von IRT-Modellen ist zudem, dass etablierte globale Kriterien zur Beurteilung der Modellgüte aktuell noch fehlen und damit die Einschätzung der absoluten Passung der Daten auf das Modell erschwert wird (Hartig & Frey, 2013).

5.2.1 DAS EINDIMENSIONALE RASCH-MODELL

Das einfachste Modell der Item Response Theory ist das Rasch-Modell. Es ist nach Georg Rasch benannt, der 1960 als erster ein probabilistisches Testmodell, das die logistische Funktion zur Beschreibung der Itemcharakteristik-Kurve (ICC) verwendet, formulierte. Die Modellgleichung des Rasch-Modell lautet (vgl. Rost, 2004, S. 119):

$$P(x_{vi}) = \frac{e^{(x_{vi}(\theta_v - \sigma_i))}}{1 + e^{(\theta_v - \sigma_i)}}$$

Die Modellgleichung beschreibt die Wahrscheinlichkeit einer bestimmten Antwort in Abhängigkeit ausgewählter Modellparameter. $P(x_{vi})$ gibt die Wahrscheinlichkeit an, mit der eine Person v das Item i löst. Der individuelle Fähigkeitsparameter wird als θ_v definiert. σ_i ist der Itemparameter, der den Abszissenwert der 50%-Lösungswahrscheinlichkeit und damit den Wendepunkt der ICC definiert. Die Steigung dieser logistischen Funktion ist für alle Testaufgaben gleich, d.h. je nachdem ob das Item schwieriger oder leichter ist, verschiebt sich die Kurve parallel der X-Achse nach rechts oder links. Damit wird angenommen, dass alle Items die gleiche Trennschärfe haben, wenn das Rasch-Modell für diesen Test gilt (vgl. Rost, 2004). Die Wahrscheinlichkeit einer bestimmten Antwort wird allein durch die Differenz aus Fähigkeitsparameter θ_v einer Person v und der Schwierigkeit σ_i eines Items i bestimmt. Unterschiede bestehen also allein in der Itemschwierigkeit (Moosbrugger, 2012). Damit postuliert das Rasch-Modell, dass die Lösungswahrscheinlichkeit für ein Item i sowohl von der Personenfähigkeit als auch der Aufgabenschwierigkeit abhängt (Moosbrugger, 2012). Die Personen- und Itemparameter werden im Rasch-Modell auf einer Skala, der Logit-Skala, dargestellt und können über die Wright Map direkt miteinander verglichen werden (Hambleton & Jones, 1993). Dabei gilt es zu beachten, dass die lokale stochastische Unabhängigkeit nicht verletzt wird und die Beantwortung eines Items nicht von der Bearbeitung eines anderen

Items abhängt (Wu & Adams, 2007). „Lokal“ steht in diesem Zusammenhang dafür, dass die Unabhängigkeit des Items nur dann gelten muss, wenn eine Person, oder mehrere Personen mit der gleichen Fähigkeit, betrachtet werden. Das heißt wiederum, dass es durchaus eine Person geben darf, die mit höherer Fähigkeit alle Aufgaben mit höherer Wahrscheinlichkeit löst als eine Person mit niedrigerer Fähigkeit (vgl. Strobl, 2012). Eindimensionale Modelle wie das Rasch-Modell, beinhalten zur Modellierung von Unterschieden zwischen Personen eine einzelne, kontinuierliche latente Variable (z.B. Embretson & Reise, 2000; Wilson, 2005). Damit werden Unterschiede in der zu erfassenden Kompetenz auf einem einzelnen Kontinuum beschrieben, d.h. jede erfasste Kompetenz entspricht einer einzelnen latenten Dimension (vgl. Hartig & Klieme, 2006; Adams, 2005). An der Modellgleichung erkennt man diese Annahme daran, dass Personen- und Itemparameter voneinander abgezogen werden und sowohl die Personenfähigkeiten als auch die Itemschwierigkeiten auf der X-Achse abgetragen sind (vgl. Strobl, 2012). Dies wiederum heißt, dass alle Personen und alle Aufgaben auf einem einzelnen Kontinuum von niedriger bis hoher Kompetenz bzw. niedriger bis hoher Itemschwierigkeit angeordnet werden. Dadurch ist das Rasch-Modell in seiner Aussagekraft eingeschränkt, da die Testitems so modelliert werden, dass alle die gleiche latente Personenfähigkeit abprüfen.

In der vorliegenden Untersuchung bedeutet dies, dass der Umgang mit MER durch eine Dimension abgebildet wird und damit keine Teilkomponenten unterschieden werden, sondern ausschließlich von Repräsentationskompetenz als eine Fähigkeit auszugehen ist. In diesem eindimensionalen Modell zur Integration von MER (vgl. ABBILDUNG 15) wird angenommen, dass sich bei der Text-Bild-Integration von Repräsentationen keine einzelnen Fähigkeiten differenzieren lassen. Diese Modellierung ist notwendig, um die modellkonform konstruierten Items auf ihre Eignung als Testaufgaben zu überprüfen und damit Gegenstand der Vorstudie (s. Kapitel 4.2.2.).

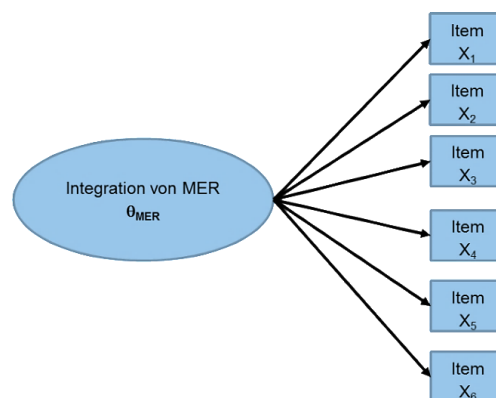


ABBILDUNG 15: 1-DIMENSIONALES MODELL ZUR INTEGRATION VON MER

5.2.2 SCHÄTZUNG VON ITEM- UND PERSONENPARAMETERN

Ein Modell ist gültig, wenn das Antwortmuster der Testpersonen zu den Annahmen des Modells passt. Mit steigender Personenfähigkeit steigt die Wahrscheinlichkeit einer korrekten Lösung des Testitems (Bühner, 2006). Neben der Personenfähigkeit ist die Lösungswahrscheinlichkeit auch von der Itemschwierigkeit abhängig. Aus diesem Grund enthalten probabilistische Testmodelle zwei Parameter: *Personenparameter* θ_i (Fähigkeitsausprägung einer Person, die aufgrund der bearbeiteten Itemanzahl einer Testperson geschätzt wird) und *Itemparameter* σ_i (Itemschwierigkeit, die auf der Lösungshäufigkeit eines Items basiert). Die Fähigkeit einer Testperson hängt von der Anzahl der gelösten Aufgaben ab, dabei ist diese Fähigkeit unabhängig davon, welche Aufgaben die Testperson im Einzelnen gelöst hat (Neumann, 2014). Hartig & Frey (2013) weisen darauf hin, dass Modelle mit Aufgabeneigenschaften nicht zu einfach formuliert werden sollten. Grund sind mögliche Zusammenhänge und Interaktionen von Aufgabeneigenschaften und Aufgabenschwierigkeiten sowie zwischen Aufgaben- und Personenmerkmalen. Das Aufgabenspektrum in einem psychometrischen Test ist damit über alle Schwierigkeitsstufen, die am Ende gemeinsam auf einer Skala aufgetragen werden, zu operationalisieren.

Die Rasch-Skalierung ist eine probabilistische Testauswertung, da die „wahre“ Fähigkeit einer Testperson mittels Schätzverfahren (z.B. dem Maximum-Likelihood-Verfahren, Rost, 2004) aus den empirischen Daten der Aufgabenbearbeitung ermittelt wird (Kauertz, 2014). Für die Modellierung der Personenfähigkeit können verschiedene Personenschätzer zu unterschiedlichen Fragestellungen herangezogen werden (Hartig & Kühnbach, 2006). Die Maximum Likelihood Estimates (MLEs) der Maximum-Likelihood-Methode sagen das gegebene Antwortmuster am wahrscheinlichsten vorher. Hartig & Kühnbach (2006) weisen darauf hin, dass MLEs messfehlerbehaftet sind und die Varianz der Personenfähigkeiten überschätzt werden. Grund hierfür ist, dass mit steigender Stichprobengröße immer auch die Anzahl der unbekannt Parameter steigt. Damit hat man keine gleichbleibende Anzahl der Parameter, was zu nicht konsistenten Schätzern führt (Strobl, 2012). Aus diesem Grund werden die Weighted Likelihood Estimates (WLEs) herangezogen, die ähnlich geschätzt werden, aber nicht zu stark streuen und die Messwerte gewichten (Hartig, Jude, & Wagner, 2008; Warm, 1989). Zudem gelten WLEs als reliabelste Schätzer für individuelle Personenwerte (vgl. Hartig & Kühnbach, 2006; Rost, 1996; Wu, 2005). Plausible Values (PVs) und Expected A Posteriori (EAPs) sind weitere Schätzer für Personenparameter, die Personenmerkmale in die Schätzung der Personenfähigkeit einbeziehen. Im Kontext der Item-Response-Theorie werden hierzu die geschätzte Varianz des zu messenden Merkmals und die Varianz der EAPs der

individuellen Merkmalsausprägungen verwendet. Als Skalenhomogenitätsmaß der Personenparameterschätzung wird die IRT-basierte EAP/PV-Reliabilität herangezogen. Sie bezieht sich auf die Messgenauigkeit von Items für ein eindimensionales Merkmal (vgl. Rost, 2004).

Bezüglich der statistischen Auswertung der Rasch-Analyse in ConQuest ist die Entscheidung zu treffen, ob der Mittelwert der Verteilung der Personenfähigkeit oder der Verteilung der Aufgabenschwierigkeit als Referenz dienen soll, da kein absoluter Bezugspunkt vorliegt (vgl. Neumann, 2014; Wu et al., 2007, Rost, 2004). Dieser kann je nach Analyse auf Null festgelegt werden. Für beide Parameter resultierten dabei üblicherweise Werte in einem Bereich zwischen -3 Logits und +3 Logits (Rauch & Hartig, 2008). Setzt man den Mittelwert der Personenfähigkeit fest und auf Null, kann z.B. das Verhalten von Aufgaben in vergleichbaren Stichproben untersucht werden. Im Rahmen der vorliegenden Studie erfolgt die Analyse zunächst auf Basis der gesamten Stichprobe, d.h. die Aufgaben werden daraufhin normiert und die Summe der Personenparameter gleich Null gesetzt. Damit wird eine sehr genaue Itemparameterschätzung (Linacre, 1994) sowie eine normorientierte Interpretation der Personenparameter möglich (Rost, 2004). Umgekehrt erreicht man mit der Normierung der Itemparameter auf Null eine kriteriumsorientierte Interpretation der Personenparameter (Rost, 2004). Itemschwierigkeiten und Personenfähigkeiten steigen von den negativen zu den positiven Werten auf der Logit-Skala hin an. Ist der Personenparameter genauso groß wie der Itemparameter, ist die Lösungswahrscheinlichkeit für dieses Item genau 50 Prozent. Items darüber sind damit schwieriger zu lösen als Items unterhalb (Boone & Scantlebury, 2006). Die Interpretation ist damit auf eine mittlere Itemschwierigkeit vorzunehmen. Bei einem Logit von -1.1 liegt die Lösungswahrscheinlichkeit für ein Item bei 25 Prozent, bei +1.1 bei 75 Prozent (Rost, 2004). Aus der Differenz zweier Personenparameter lassen sich schließlich Aussagen über den Fähigkeitsunterschied zweier Personen machen, die unabhängig von der Normierung der Itemparameter und von Fähigkeitsausprägungen anderer Testpersonen sind (Rost, 2004, S. 121). Die Normierung wird dabei unabhängig für ein- und mehrdimensionale Rasch-Modelle vorgenommen.

Die Genauigkeit einer Schätzung der Itemschwierigkeit ist abhängig von der Anzahl der Antworten pro Item. In der Literatur finden sich bei Linacre (1994) Richtwerte, die Hinweise auf die zu erwartende Genauigkeit der geschätzten Itemschwierigkeit geben. Ausgehend von einem Konfidenzintervall von 95% liegt die Genauigkeit der geschätzten Itemschwierigkeit bei einer Häufigkeit von mindestens 30 Antworten pro Item bei $\pm 1,0$ Logits. Liegen 100 Antworten pro Item vor, wird die Genauigkeit der Schätzung auf $\pm 0,5$ Logits erhöht. Dieser Wert liegt ebenfalls vor, wenn jedes Item von 100 Testpersonen beantwortet wird und ein Kon-

fidenzintervall von 99% erreicht werden soll. Die Genauigkeit einer Schätzung der Personenparameter ist abhängig von der vorgelegten Anzahl an Items pro Testperson (Rost, 2004). Die Itemzahl kann dabei eine direkte Auswirkung auf die Reliabilität haben, wenn den Testpersonen nur eine begrenzte Anzahl an Testaufgaben zur Bearbeitung gegeben wird (Ropohl, 2010; Schmiemann, 2010; Senkbeil et al., 2005; Wellnitz, 2012). Nach Grube (2011) konnten zufriedenstellende Personenreliabilitäten bei einer Beantwortung von sechs Items pro Testperson erreicht werden.

5.3 AUSWERTUNGSMETHODEN DER VORSTUDIE

5.3.1 VORBEREITENDE SCHRITTE

Zur Beantwortung der Testaufgaben wurde eine itemspezifische Kodierung vorgenommen (Rost, 2004). Dabei wurden folgende Codes für die Beantwortung der Kategorien festgelegt (vgl. TABELLE 5).

TABELLE 5: CODES FÜR DIE ITEMSPEZIFISCHE KODIERUNG

Code	Definition
99	Testaufgabe/ Item wurde nicht beantwortet
79	Testaufgabe war nicht im Testheft enthalten
1	Beantwortung der festgelegten Kategorie ist richtig
0	Beantwortung der festgelegten Kategorie ist falsch

Die Items der vorliegenden Arbeit sind so konstruiert, dass alle Antworten der Testpersonen als richtig oder falsch codiert werden. Damit liegt ein dichotomer Leistungstest vor, d.h. unabhängig davon, ob ein geschlossenes, halboffenes oder offenes Antwortformat vorliegt, wird die Bearbeitung einer Testaufgabe eindeutig als falsch (0, no credit) oder richtig (1, full credit) bewertet.

5.3.2 QUALITATIVE INHALTSANALYSE

Im Rahmen einer Masterarbeit (Schechert, unveröffentlicht, 2014) wurden alle Testaufgaben der Testheftgruppe A aus dem Aufgabenkontext Ökologie mit Methoden der deskriptiven Statistik analysiert. Anhand induktiver und deduktiver Kriterien (vgl. Mayring, 2000) wurde ein elaboriertes Kategoriensystem zur Deutung des Aufgabenmaterials erstellt (vgl. Bortz & Döring, 2006). Die Aufgaben im offenen Antwortformat wurden dementsprechend codiert und

für die spätere Selektion einzelner Testitems ausgewertet. Auf Basis einschlägiger Literatur und bestehender Theorien zum Umgang mit MER wurden deduktive Kategorien entwickelt, die die Integration von MER definieren sollen. Neben dem Strukturmodell zur Diagrammkompetenz (Lachmayer, 2008), dem Konzept der *Representational Competence* (Kozma & Russell, 2005) sowie dem Kompetenzmodell von Klöpfel, Schwanewedel, & Mayer (2013) zum Erschließen biologischer Informationen aus unterschiedlichen Repräsentationen (Text und Bild) flossen zusätzlich die Ergebnisse der qualitativen Studie von Anderson et al. (2013) zum Umgang mit Schemazeichnungen in die deduktive Kategorienbildung ein. Dabei wurden weiterhin die Schwierigkeitsstufen gemäß der Einheitlichen Prüfungsanforderungen für das Abitur (KMK, 2004) sowie die vier Typisierungen biologischen Wissens nach Schönborn & Bögeholz (2009), Begriffe – Konzepte – Prinzipien – Grundsätze, in Anlehnung an die fünf Komplexitätsstufen nach Kauertz et al. (2010) berücksichtigt. Neben der Formulierung von Oberkategorien erfolgte anhand der zusammenfassenden Inhaltsanalyse (vgl. Mayring, 2010) die induktive Bildung von Unterkategorien. Dabei wurde das vollständige Datenmaterial analysiert und auf explizite Teilfähigkeiten der Testpersonen reduziert. Zusätzlich wurde beachtet, dass die Kategorien logisch voneinander unabhängig sind (vgl. Rost, 2004). Damit konnte ein strukturiertes, elaboriertes Kategoriensystem entwickelt werden, das nicht zu feingliedrig, nicht zu umfangreich, dennoch trennscharf und in Beziehung zu den Fragestellungen und Zielsetzungen der Untersuchung steht.

Für die präzise Zuordnung der Probandenlösungen zu den Kategorien wurde ein Kodierleitfaden entwickelt. Anhand trennscharfer Definitionen und Ankerbeispielen wird die Qualität der Kodierung durch ein höheres Übereinstimmungsergebnis unterschiedlicher Kodierer verbessert (Hammann & Jördens, 2014; Bortz & Döring, 2006). Für die weiterführenden Analysen im Rahmen der Klassischen und Probabilistischen Testtheorie wurden die beantworteten offenen Aufgaben der Testpersonen anhand des entwickelten Kodierleitfadens codiert. Das Kodiermanual wurde vor Durchführung der Pilotstudie entwickelt, an kleineren Stichproben erprobt, überarbeitet und finalisiert (vgl. Hammann & Jördens, 2014). Für die sich anschließende Kodierung wurden die Kategorien auf ihre Trennschärfe geprüft. Trennscharfe Kategorien wurden 0/1-codiert. Die finale Codierung des offenen Items wurde über die Aggregation der Kategorienwerte ermittelt. Anhand dieser Summenwerte für jede einzelne Testaufgabe (Summe der richtig codierten Kategorien pro Testitem) wurde je nach Schwierigkeit der Kategorien ein Cut-Off-Wert zwischen 50 und 65% festgelegt und ab hier mit dem Code 1 bewertet. Dabei wurde darauf geachtet, dass Kategorien, die über alle Aufgaben keine Informationsaufklärung anhand von Unterschieden im Beantwortungsmuster zeigen, ausgeschlossen und nicht zur Berechnung des Summenwerts herangezogen wurden. Damit wird

sichergestellt, dass nur diejenigen Kategorien eine Bewertung erhalten, die auch tatsächlich Differenzwerte in der Bearbeitung der Testpersonen zeigen. Im Rahmen der Kodierung innerhalb der Vorstudie wurde die inhaltliche Beschreibung der Codes sukzessive angepasst und optimiert. Damit liegt ein dichotomer Leistungstest vor, d.h. unabhängig davon, ob ein geschlossenes, halboffenes oder offenes Antwortformat vorliegt, wird die Bearbeitung einer Testaufgabe eindeutig als falsch (0, no credit) oder richtig (1, full credit) bewertet. Dieses Kodierschema wurde für die Hauptstudie ebenfalls übernommen, hingegen erfolgte die Kodierung direkt im Full- und No Credit-Verfahren (s. Kapitel 7).

Vor dem Hintergrund der drei Testgütekriterien Objektivität, Reliabilität und Validität wurden die Schritte der qualitativen Inhaltsanalyse stets dokumentiert und können als gewährleistet angenommen werden. Das methodische Vorgehen wurde möglichst genau beschrieben, transparent formuliert und detailliert dargestellt, um das Kriterium der Objektivität zu sichern (vgl. Bortz & Döring, 2006). Diese Schritte wurden von zwei Personen unabhängig durchgeführt, die Ergebnisse verglichen (vgl. Krüger & Riemeier, 2014) und damit die Auswertungs- und Interpretationsobjektivität eingeschätzt (vgl. Bortz & Döring, 2006). Da die Methoden der quantitativen Forschung zur Reliabilitätsmessung nicht ohne Weiteres auf qualitative Untersuchungen übertragen werden können, wird zur Zuverlässigkeitsprüfung die Intrarater-Reliabilität als Gütekriterium herangezogen (vgl. Bortz & Döring, 2006; Krüger & Riemeier, 2014). Bei diesem Vorgehen werden etwa zehn Prozent des Datenmaterials vom gleichen Kodierer zu einem späteren Zeitpunkt erneut analysiert und geprüft, ob die Ergebnisse reproduziert werden können (vgl. Krüger & Riemer, 2014; Hammann & Jördens, 2014). Der Interrater-Reliabilitätskoeffizient wird hingegen als Maß bezeichnet, „mit dem mehrere Personen, die unabhängige Kodierungen vornehmen, offene Aufgaben denselben Antwortkategorien zuordnen“ (Hammann & Jördens, 2014). Damit wird der interpersonale Konsens beschrieben, indem unterschiedliche Kodierer bei der gleichen Untersuchung zu vergleichbaren Ergebnissen kommen (Bortz & Döring, 2006). Anstelle der Berechnung einer prozentualen Beobachterübereinstimmung wird als statistisches Maß der Cohens Kappa Koeffizient κ verwendet, der sich folgendermaßen berechnen lässt (vgl. Grouven et al., 2007):

$$\kappa = \frac{\rho_0 - \rho_e}{1 - \rho_e}$$

Dabei steht ρ_0 für den Anteil tatsächlich beobachteter Übereinstimmungen von Rater A und Rater B und ρ_e für den Anteil zufälliger Übereinstimmungen. Ein Cohens Kappa von 1 bedeutet, dass eine vollständige Übereinstimmung zwischen beiden Kodierern vorliegt. Greve & Wentura (1997) sprechen ab einem Cohens Kappa über 0,75 von einer guten bis ausgezeichneten Interrater-Reliabilität. Bortz & Döring (2006) sehen Werte zwischen 0,60 und 0,75 als

gute Übereinstimmungswerte. Zur Bestimmung der Interrater-Reliabilität werden ebenfalls ca. 10% der kodierten Aufgaben aus der behandelten Stichprobe durch einen zweiten Codierer (= Rater) erneut kodiert und der Cohens Kappa ermittelt. Liegen die Werte über den genannten Kennzahlen kann die Codierung als ein reliables Verfahren bewertet werden, das das Lösungsverhalten der Testpersonen zum Erwartungshorizont korrekt interpretiert. Über die Validität wird ermittelt, inwieweit der entwickelte Test das intendierte Konstrukt erfasst. Die Validität gilt als entscheidendes Kriterium der qualitativen Inhaltsanalyse (Mayring, 2010) und erhebt den Anspruch, dass die Inhaltsanalyse durch eine sorgfältige Verfahrensdokumentation nachvollziehbar sein muss. Hierzu zählen die semantische Validität, Triangulation und die kommunikative Validierung.

Im Rahmen der vorliegenden Arbeit wird die Übereinstimmung über das Datenmaterial der offenen Aufgaben anhand einer einfachen Übereinstimmungsmatrix (vgl. Rost, 2004; Bortz & Döring, 2006) in IBM SPSS 22 errechnet. Als Richtwerte zur Interpretation der Interrater-Reliabilität können für den Koeffizienten κ die in der TABELLE 6 zu entnehmenden Intervalle unterschieden werden (vgl. Altmann, 1991; Landis & Koch, 1977):

TABELLE 6: RICHTWERT FÜR COHENS KAPPA KOEFFIZIENT

Wert von Cohens Kappa κ	Stärke der Übereinstimmung
< 0,20	schwach (slight)
0,21 – 0,40	leicht (fair)
0,41 – 0,60	mittelmäßig (moderate)
0,61 – 0,80	gut (substantial)
0,81 – 1,00	sehr gut (almost perfect)

Die Intrarater-Reliabilität wurde von Rater 1 nach zwei Wochen erneut bestimmt und der Cohens Kappa Koeffizient berechnet ($\kappa = 0,798$). Laut TABELLE 6 kann bei diesem Wert von einer guten Übereinstimmung in der Codierung ausgegangen werden. Der Wert resultiert aus folgender Berechnungsmatrix (vgl. TABELLE 7).

TABELLE 7: KREUZTABELLE ZUR BERECHNUNG DER INTRARATER-RELIABILITÄT

		Rating 2		Gesamtsumme
		richtig	falsch	
Rating 1	richtig	89	9	98
	falsch	4	42	46
Gesamtsumme		93	51	144

Die Interrater-Reliabilität wurde ebenfalls bestimmt und anhand von etwa zehn Prozent der codierten Aufgaben aus der behandelten Stichprobe durch einen zweiten Rater erneut codiert. Zur Berechnung wurden 174 Variablen (bzw. Kategorien) herangezogen und der Cohens Kappa Koeffizient bestimmt ($\kappa = 0,804$). Als Grundlage diente erneut eine Berechnungsmatrix (vgl. TABELLE 8).

TABELLE 8: KREUZTABELLE ZUR BERECHNUNG DER INTERRATER-RELIABILITÄT

		Rater/ Kodierer 1		Gesamtsumme
		richtig	falsch	
Rater/ Kodierer 2	richtig	38	7	45
	falsch	6	123	129
Gesamtsumme		44	130	174

Auch in diesem Fall kann von einer guten bis sehr guten Übereinstimmung zwischen den beiden Ratern ausgegangen werden. Im Rahmen der Hauptstudie wurden sowohl die Intrarater- als auch die Interrater-Reliabilität erneut, nach den beschriebenen Kriterien, bestimmt. Dabei resultierte in beiden Fällen eine sehr gute Übereinstimmung. Die Intrarater-Reliabilität liegt bei einem Cohens Kappa von $\kappa = 0,936$ (Gesamtsumme $n=135$ Fälle) und die Interrater-Reliabilität bei einem Cohens Kappa von $\kappa = 0,826$ (Gesamtsumme $n=72$ Fälle).

5.3.3 ANALYSE DER ITEMPARAMETER

Bei der Rasch-Analyse stehen verschiedene Statistiken zur Überprüfung der Testaufgaben auf Rasch-Homogenität zur Verfügung. Das vorliegende Kapitel beschreibt Merkmale, die zum Verständnis der Itemerprobung und der anschließenden Itemanalyse sowie der Itemselektion dienen. Die Pilotierung der vorliegenden Untersuchung hatte eine empirische Validierung der entwickelten Testaufgaben hinsichtlich der Konstruktionsanleitung zum Ziel und wird über psychometrische Kennwerte dokumentiert. Zusätzlich werden die Testaufgaben aus fachdidaktischer Expertise begutachtet. Damit können Aufgaben für die eigentliche Untersuchung, die Hauptstudie, ausgewählt werden.

Bei der Evaluation der Items werden psychometrische Kennwerte herangezogen, um damit zu überprüfen, ob die getroffenen Modellannahmen stimmen und die Items tatsächlich modellkonform sind (Moosbrugger, 2012). Für jede Testaufgabe werden der Schätzer (Estimate) der Aufgabenschwierigkeit sowie der entsprechende Fehler angegeben. Aufgrund der empirischen Beschreibung von Repräsentationskompetenz im Rahmen der Item Response Theory (IRT) werden daraufhin folgende charakteristische IRT-Merkmale betrachtet:

- die Itemschwierigkeit (Rost, 2004),
- die Item-Characteristic-Curve (ICC) (Wu & Adams, 2007) und der
- Item-Fit-Maße: Mean-Square-(MNSQ)-Fit-Statistiken (Wu & Adams, 2007) sowie T-Werte (Bond & Fox, 2007; Neumann, 2014).

Für die erste quantitative Beurteilung der Testaufgaben wird die *Itemschwierigkeit* (σ_i) als deskriptives Maß herangezogen. Items, die sich als extrem leicht oder extrem schwer zu bearbeiten erweisen, müssen aus dem Test ausgeschlossen werden (z.B. Kelava & Moosbrugger, 2012; Terzer et al., 2013). Im Rahmen der probabilistischen Testtheorie bezeichnet man als Itemschwierigkeit denjenigen Fähigkeitswert, der einer Wahrscheinlichkeit einer korrekten Lösung von 0,5 entspricht (Rost, 2004). Dabei entspricht eine Itemschwierigkeit von z.B. 2,0, dass Personen mit einem Fähigkeitswert über 2,0 das Item mit einer Wahrscheinlichkeit von über 50% korrekt lösen können. Der Zusammenhang zwischen Personenfähigkeit und der Wahrscheinlichkeit einer bestimmten Itemantwort kann in Form einer *Itemcharakteristik-Kurve (ICC)* abgebildet werden. Aufgetragen wird $P(x_{vi}=1)$, d.h. die Wahrscheinlich für eine korrekte Antwort, codiert als 1, in Abhängigkeit von der Personenfähigkeit θ_v (vgl. ABBILDUNG 16).

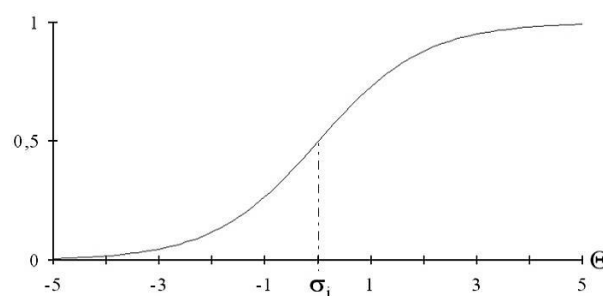


ABBILDUNG 16: DIE ITEMCHARAKTERISTIK-KURVE (ICC) DES RASCH-MODELLS FÜR EIN ITEM MIT DER ITEMSCHWIERIGKEIT $\sigma_i=0$

Die ICC beschreibt die Lösungswahrscheinlichkeit eines Items für unterschiedliche Personenfähigkeiten (Terzer, 2012). Für die Modellierung von Fähigkeiten, die zum Lösen von Aufgaben eines bestimmten Typs benötigt werden, werden Modelle gewählt, die eine monoton

steigende Funktion annehmen. In diesem Fall steigt die Kurve mit steigender Personenfähigkeit an und zeigt einen sigmoiden Verlauf. Zur Analyse von Items sollte die empirische ICC möglichst nah an der theoretisch vorhergesagten Kurve liegen (Wu & Adams, 2007). Damit kann eine optische Kontrolle der Items zwischen angenommenen Verlauf der Lösungswahrscheinlichkeit (ICC) und dem empirischen Verlauf der Lösungswahrscheinlichkeit der Items detektiert werden (Bond & Fox, 2007; Embretson & Reise, 2000). Zu überprüfen gilt es, ob die theoretische und die empirische Kurve bei einem „guten“ Item nahe beieinanderliegen (Wu & Adams, 2007). Die Lage auf der Fähigkeitsskala ist dabei ein Maß für die Itemschwierigkeit. Je weiter rechts die Kurve liegt, desto schwieriger ist das entsprechende Item. Dies wiederum bedeutet, dass nur Personen mit einem bestimmten Fähigkeitswert das Item mit einer Lösungswahrscheinlichkeit von mindestens 50% beantworten können.

Neben der Itemschwierigkeit wird für die Itembeurteilung die *Trennschärfe* (r_{it}), d.h. die Übereinstimmung der Differenzierung eines einzelnen Items mit der Differenzierung aller übrigen Items innerhalb des Tests, betrachtet (z.B. Amelang & Zielinski, 2002; Kevala & Moosbrugger, 2012). Dabei ist die Trennschärfe ein Maß dafür, wie gut sich Personen verschiedener Fähigkeiten unterscheiden lassen (Rost, 2004). Die Trennschärfe wird in der klassischen Testtheorie als punkt-biseriale Korrelation zwischen Itembeantwortung und dem Gesamtestwert berechnet (vgl. Kelava & Moosbrugger, 2007; Bortz, 2005; Fisseni, 1997) und hat einen korrelationstypischen Wertebereich von -1,0 bis +1,0. Der Signifikanztest für die punkt-biseriale Korrelation entspricht dem t-Test für unabhängige Stichproben (Bortz & Schuster, 2010). Dabei gelten Werte größer als 0,5 als hoch (Bortz & Döring, 2006). Items, die eine negative Korrelation mit dem Gesamtestwert zeigen sowie Items mit einer Trennschärfe kleiner 0,25 sollten aus dem Test entfernt werden (Kelava & Moosbrugger, 2012; Adams & Wu, 2002). Wie bereits erwähnt, steht die Trennschärfe in Zusammenhang mit dem Schwierigkeitsindex. Je höher die Schwierigkeit eines Items, desto geringer ist die Trennschärfe (Bortz & Döring, 2006). In der probabilistischen Testtheorie wird die Steigung, d.h. der Bereich mit dem steilsten Anstieg als Tangente, der Itemcharakteristik-Kurve eines Items als Maß für die Trennschärfe herangezogen. Dabei gilt, je flacher die ICC, desto weniger trennscharf ist das Item. Je steiler die Kurve, desto trennschärfer sind die Items. Abschließend gilt es zu beachten, dass bei Vorliegen einer mehrdimensionalen Modellierung des Tests, wie im Rahmen der vorliegenden Untersuchung, die Trennschärfe getrennt für die Items je Teilkompetenz betrachtet werden muss. Zusätzlich werden der Aufgaben-Fit (*Item-Fit*) zum Rasch-Modell sowie die Einschränkung, dass die Anforderungen nicht in einem extremen Randbereich der Skala liegen, überprüft (vgl. Kauertz, 2014).

Der *Item-Fit* wird herangezogen, wenn es um die Passung eines Items zum verwendeten Messmodell, hier ein- bzw. mehrdimensionale IRT-Modelle, geht. Die vorhergesagten Lösungshäufigkeiten eines Items werden mit den tatsächlich beobachteten verglichen. Als Referenzwert wird der gewichtete („*weighted mean square*“, wMNSQ) Item-Fit, auch *In-Fit* bezeichnet, verwendet. Dieser beruht auf einer Analyse der Abweichung der beobachteten Aufgabebearbeitung durch die Personen von dem auf Basis des Rasch-Modells erwarteten Werts (Neumann, 2014), d.h. der wMNSQ ist ein Maß für die Abweichung zwischen theoretisch angenommen Verlauf der Lösungswahrscheinlichkeit und dem empirisch beobachteten Verlauf (Bond & Fox, 2007). Der wMNSQ kann einen Erwartungswert von 1,0 annehmen und gibt an, inwiefern die Steigung, d.h. die Trennschärfe eines Items, vom Messmodell vorhergesagt wird. Er berechnet sich als Quotient aus dem Mittelwert der quadrierten beobachteten Abweichungen und dem Mittelwert der erwarteten Abweichungen (vgl. Wright & Stone, 1999). Ein idealer Wert ist 1,0, Abweichungen im Bereich von 0,8 bis 1,2 werden als akzeptabel erachtet und die Items gelten als raschkonform (Bond & Fox, 2007; Wright & Linacre, 1994). Sind die Werte größer als 1,0, resultiert eine flachere ICC und damit eine geringere Trennschärfe des Items als vorhergesagt wurde. Werte unter 1 führen zu einer steileren ICC und damit eine größere Trennschärfe. Aus diesem Grund können Items mit einem wMNSQ unter 1,0 eher im Itempool belassen werden (Wu & Adams, 2007). Der wMNSQ ist stichprobenabhängig und geht bei ausreichender Testpersonenanzahl gegen 1,0. Aus diesem Grund wird zusätzlich eine inferenzstatistische Prüfung durch den T-Wert durchgeführt. Der T-Wert resultiert aus der Transformation in eine T-Verteilung unter Berücksichtigung der Stichprobe und kann von ConQuest ebenfalls ausgegeben werden. Die T-Verteilung hat einen Mittelwert von 0 und eine Standardabweichung von 1 und erlaubt eine Einschätzung über die Diskriminierung der Items vom Modell, d.h. inwieweit der zugehörige MNSQ-Wert vom Erwartungswert 1 abweicht (Wu & Adams, 2007). Sind die T-Werte $> 2,0$ oder $< -2,0$ wird damit ein signifikant zu guter bzw. zu schlechter Fit signalisiert (Bond & Fox, 2007). Rost (2004) schlägt einen Wertebereich zwischen -1,96 und + 1,96 vor. Für die Überprüfung der Testaufgaben müssen je nach Untersuchungszweck die Items anhand unterschiedlicher Kriterien analysiert werden.

Für die Itemanalyse werden zunächst eindimensionale Raschskalierungen durchgeführt. Zur Beurteilung der Modellpassung werden einzelne Modellparameter (Item-Fits), die gewichteten Abweichungsquadrate (wMNSQ) und die dazugehörigen T-Werte herangezogen. Bezüglich der wMNSQ-Werte werden nur Items mit Werten $> 2,0$ entfernt. Items mit wMNSQ-Werten $< 1,0$ bleiben im Test. Für die T-Werte gilt dies ebenfalls, T-Werte von $T < -2,0$ werden nicht entfernt, da „even an item with a very low mean-square tells us a little something that

is new and useful“ (Wright & Linacre, 1994, S. 370). Des Weiteren erfolgte die Entscheidung über Elimination oder Verbleib von Testitems im Gesamtpool, wie bereits beschrieben, unter Berücksichtigung der klassischen Trennschärfe sowie einer inhaltlichen und fachdidaktischen Begutachtung.

5.3.4 DESKRIPTIVE ANALYSEN IM RAHMEN DER KLASSISCHEN TESTTHEORIE

Im Rahmen der klassischen Testtheorie werden die in den Rasch-Skalierungen geschätzten Item- und Personenparameter weiter analysiert und in SPSS überführt. In diesem Zusammenhang kommen multivariate Analysemethoden zum Einsatz. Die Verfahren der multivariaten Analyse werden in IBM SPSS 22 durchgeführt.

Für die Datenauswertung stehen je nach Skalenniveau und Datenverteilung parametrische oder nicht-parametrische (verteilungsfreie) Verfahren zur Verfügung. Für parametrische Verfahren müssen bestimmte Voraussetzung für die abhängige Variable gegeben sein wie mindestens intervallskalierte Daten, Normalverteilung, Varianzhomogenität und stochastische Unabhängigkeit (Bühner & Ziegler, 2009; Field 2009). Die erste Voraussetzung kann als erfüllt angesehen werden, da die geschätzten abhängigen Variablen, Item- und Personenparameter, ein metrisches Skalenniveau besitzen. Die Prüfung der empirischen Daten auf Normalverteilung erfolgt mit dem Shapiro-Wilk-Test (Field, 2009). Wird der Test signifikant ($p < 0,05$), liegt eine signifikante Abweichung von der Normalverteilung vor. Die Varianzhomogenität, d.h. die homogene Verteilung einer Variablen in der Grundgesamtheit, wird mit dem Levene-Test überprüft (Bühner & Ziegler; 2009; Field, 2009). Fällt dieser signifikant ($p < 0,05$) aus, gilt die Varianzhomogenität als nicht angenommen. Mithilfe des Korrelationskoeffizienten kann die stochastische Unabhängigkeit ermittelt werden, die immer dann gegeben ist, wenn die Korrelationen der Mittelwerte verschiedener Gruppen $r < 0,05$ ist. In diesem Fall trifft die Nullhypothese, d.h. dass die Mittelwerte der Gruppen gleich schwer sind, für die stochastische Unabhängigkeit zu (Field, 2009). Werden diese oder weitere Voraussetzungen für die parametrischen Verfahren nicht erfüllt oder liegen ordinale Daten vor, kommen nicht-parametrische Verfahren zum Einsatz. Je nachdem ob die Daten normalverteilt, die Varianzen homogen oder die Stichprobenumfänge gleich oder unterschiedlich groß sind, muss der passende Test ausgewählt werden (Bortz, 2005).

Verhältnis der Schwierigkeiten: Für die Strukturbeschreibung der Repräsentationskompetenz im Rahmen der Integration von MER wird mittels deskriptiver Analysen die Itemschwierigkeit analysiert. Für die Darstellung der Ergebnisse wird untersucht, in welchem Verhältnis die Schwierigkeiten der (Teil-) Komponenten zueinanderstehen. Dabei erfolgt ein Vergleich

der rasch-skalierten Itemparameter (vgl. Einhaus, 2007). Für die graphische Darstellung werden Boxplots verwendet, die als abhängige Variable die Itemschwierigkeit auftragen, während die unabhängigen Variablen die (Sub-) Komponenten auf der X-Achse angeben. Die Werte der Itemschwierigkeit gehen von -3 bis +3, wobei leichtere Items einen geringeren Wert haben. Als zentraler Wert wird der Median eingetragen, als obere und untere Grenze einer „Box“ ist das 75. und 25. Perzentil aufgetragen. Dies bedeutet, dass innerhalb der Box 50 Prozent der Messwerte der Items der jeweiligen Komponente liegen (Döring & Bortz, 2016). Die „Whisker“ zeigen die Spannweite der Verteilung an und damit die Werte vom Minimum bis zum Maximum (Bühner & Ziegler, 2009; Dytham, 2005). Um Ausreißer- und Extremwerte zu bestimmen wird der Interquartilabstand herangezogen und in SPSS bei Vorliegen eines Ausreißerwertes (Messwerte, die von den Boxgrenzen mehr als 1,5-mal die Kantenlänge entfernt sind) ein Kreis sowie bei Extremwerten (Abstand von mehr als 3-mal der Kantenlänge) ein Stern, jeweils versehen mit der Versuchspersonennummer, abgebildet (Brosius, 2013; Bühner & Ziegler, 2009). Für die Berechnung der Unterschiede zwischen den einzelnen (Sub-) Komponenten bezüglich der mittleren Itemschwierigkeiten wird der nicht-parametrische Kruskal-Wallis-Test für unabhängige Stichproben verwendet, da durch die unterschiedliche Itemverteilung in den (Sub-) Komponenten keine Varianzhomogenität und Normalverteilung vorliegt (vgl. Bühner & Ziegler, 2009).

Analyse von Mittelwertsunterschieden: Liegen mindestens intervallskalierte Daten vor, sind die Werte der jeweiligen Stichproben normalverteilt und Varianzhomogenität sowie stochastische Unabhängigkeit gegeben, können für den Mittelwertsvergleich *einfaktorielle Varianzanalysen (ANOVA)* sowie für den Vergleich von zwei Gruppen *t-Tests* eingesetzt werden (vgl. Field, 2009). Beide Analyseverfahren eignen sich für die Testung und Vorhersage möglicher Auswirkungen einer unabhängigen Variable (z.B. Integration von MER) auf eine abhängige Variable (z.B. Itemschwierigkeit) und sind sowohl bei Between-Designs als auch bei Within-Designs anwendbar.

Mithilfe des t-Tests für *unabhängige Stichproben* (Testpersonen bzw. Testgruppen sind unterschiedliche Personen) lassen sich Unterschiede zwischen zwei Mittelwerten ermitteln, wobei die Beobachtungen in der einen Stichprobe in keinerlei Beziehung zu den Beobachtungen der anderen Stichprobe stehen (Field, 2009; Bortz & Schuster, 2010). Mithilfe des t-Tests für *abhängige Stichproben* werden zwei Stichproben miteinander verglichen, deren Elemente einander paarweise zugeordnet sind, d.h. man geht von einer Stichprobe von Paaren aus (Bortz & Schuster, 2010). Im Rahmen der vorliegenden Untersuchung lässt sich so klären, ob aufgrund der vorliegenden Daten davon auszugehen ist, dass Items des Kompetenzbereich

MERI schwieriger sind, als Items des Kompetenzbereich MERII oder MERIII (t-Test für unabhängige Stichproben). Zusätzlich wird untersucht, ob die Kompetenzausprägung in zwei Teilkompetenzen unterschiedlich ist (t-Test für abhängige Stichproben). Statistisch signifikant werden diese Unterschiede bei einem p-value von $p < 0,05$. Bei Verletzung der Annahmen für den t-Test für unabhängige Stichproben, stochastische Unabhängigkeit und Varianzhomogenität, oder bei Vorliegen von ordinalen Daten wird der parameterfreie *Mann-Whitney-U-Test* herangezogen (Field, 2009; Bortz, 2005). Bei abhängigen Stichproben kann der Wilcoxon signed-rank Test herangezogen werden (Field, 2009). Bei Between-Designs wird die abhängige Variable (z.B. Itemschwierigkeit) genau einer von mehreren Gruppen zugeordnet und getestet. Damit können beispielsweise zuvor definierte Stufen innerhalb einer Dimension zugeteilt und getestet werden (vgl. Wellnitz, 2012). Die Berechnung der Mittelwerte der unabhängigen Gruppen erfolgt mittels einfaktorieller Varianzanalyse, liegen nur zwei Stichproben vor, wird ein t-Test für unabhängige Stichproben durchgeführt (Field, 2009). Bei Within-Designs werden t-Tests für verbundene Stichproben herangezogen, wenn jeweils zwei Mittelwerte direkt miteinander verglichen werden und die oben genannten Voraussetzungen für t-Tests erfüllt sind. Diese Analysen sollten hingegen nur dann durchgeführt werden, wenn die Differenzen zwischen den Item- und Personenkenwerten der jeweiligen Stichproben normalverteilt sind (Field, 2009).

Für den Vergleich von mehr als zwei Mittelwerten werden Varianzanalysen durchgeführt. Dabei wird überprüft, ob mehrere Mittelwerte zu einem gleichen gemeinsamen Mittelwert angehören (Field, 2009). Damit ermöglicht die ANOVA Aussagen darüber, ob sich die Mittelwerte einzelner Variablen in mehreren Fallgruppen signifikant unterscheiden. Durch eine zunächst optische Prüfung werden die gewonnenen Daten auf Normalverteilung getestet (vgl. Sedlmeier & Renkewitz, 2007). Weitere Tests werden nicht herangezogen, da die in der Varianzanalyse gewonnene F-Statistik relativ robust gegenüber einer Verletzung der Normalverteilung ist (Glass, Peckham, & Sanders, 1972). Als weitere Voraussetzung werden die Daten auf Varianzhomogenität (Levene-Test) geprüft, wobei auch hier die ANOVA robust bei Verletzungen ist (vgl. Field, 2009, S. 360). Mit der einfaktoriellen ANOVA wird im Rahmen der vorliegenden Untersuchung ein Vergleich aller Gruppenmittelwerte der Itemschwierigkeiten zu den drei Teilkompetenzen der Kompetenz *Integration von MER* durchgeführt, um mögliche Komplexitätsstufen zu identifizieren oder auszuschließen. Damit kann überprüft werden, ob Personen, die beispielsweise unterschiedliche Studiengänge besuchen, sich in Hinblick auf die mittleren Kompetenzen zur *Integration von MER* unterscheiden.

Effektstärke: Beobachtete Effekte können durch die Berechnung von Effektstärken weiter klassifiziert und interpretiert werden. Bei der Effektstärke handelt es sich um standardisierte Mittelwertsunterschiede zwischen zwei untersuchten Gruppen (Bühner & Ziegler, 2009). Für bilaterale Gruppenvergleiche werden Cohens d und für ein- und mehrfaktorielle Varianzanalysen das partielle eta-Quadrat (η^2_p) herangezogen. Dabei entspricht $d \geq 0,20$ einem kleinen, $d \geq 0,50$ einem mittleren und $d \geq 0,80$ einem großen Effekt (Cohen, 1988, zitiert nach Bühner & Ziegler, 2009, S. 177). Für η^2_p gilt: $\eta^2_p \geq 0,01$ deutet auf einen kleinen, $\eta^2_p \geq 0,06$ auf einen mittleren und $\eta^2_p \geq 0,14$ auf einen starken Effekt hin (Bühner, 2011, S. 268) bzw. unterscheiden sich kleine $\eta^2_p \geq 0,01$, mittlere $\eta^2_p \geq 0,09$ und große $\eta^2_p \geq 0,25$ Effekte (Field, 2009).

5.4 GÜTEKRITERIEN

Diagnostische Verfahren unterliegen einer Reihe von Qualitätsansprüchen. Grundlage für Generalisierungen und empirischer Evidenz sind Gütekriterien, die erst die Interpretation und Erklärung wissenschaftlicher Fragestellungen erlauben. Für die Diagnostik von Kompetenzen werden die Gütekriterien Objektivität, Reliabilität und Validität herangezogen und im Rahmen der vorliegenden Arbeit berücksichtigt.

Objektivität ist immer dann gegeben, wenn ein standardisiertes Vorgehen, eine sorgfältige Dokumentation aller Schritte vorgenommen und die Testdurchführung durch einen qualifizierten Testleiter absolviert wird (Hartig & Jude, 2007). Entsprechend kann zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität eines Tests unterschieden werden. Für die vorliegende Arbeit erfährt die Objektivität einen großen Stellenwert, da es um das Beurteilen offener Antwortformate geht und hier Verhaltensweisen durch einen Beurteiler geschätzt werden (Hartig & Jude, 2007).

Reliabilität bezeichnet die Messgenauigkeit eines Tests. Die Reliabilitätsbestimmung basiert auf der internen Konsistenz. In der KTT werden statistische Kennwerte wie Cronbachs Alpha herangezogen, während in der PTT die Schätzung der Reliabilität über Zusammenhänge der Items erfolgt (Hartig & Jude, 2007). Für Kompetenzkonstrukte mit mehrdimensionaler Struktur, wie sie auch in dieser Arbeit angenommen wird, können Schätzungen der Reliabilität alternativ über Varianzaufklärungen innerhalb der Lösungswahrscheinlichkeit erfolgen. Dabei gilt es zu beachten, dass die Reliabilität eines Tests in engem Zusammenhang zur Trennschärfe steht (Terzer et al., 2013). Eine Skala mit korrelierenden Testaufgaben führt entsprechend zu einer hohen internen Konsistenz.

Validität ist definiert als das Ausmaß, in dem ein Test das misst, was er zu messen vorgibt (Rost, 2004). Die Validität ist auch „die notwendige Legitimation für das Ziehen individueller und institutioneller Konsequenzen aus einer Testanwendung“ (Hartig & Jude, 2007, S. 22).

Die Validierung eines Tests kann durch drei Aspekte vorgenommen werden (vgl. Schmie-
mann & Lücken, 2014; Hartig & Jude, 2007; Hartig et al., 2008):

- Inhaltsvalidität, d.h. inwieweit die Testinhalte das zu messende Konstrukt erfassen.
- Kriteriumsvalidität, die überprüft, ob es einen Zusammenhang zwischen den indivi-
duellen Ausprägungen eines oder mehrerer externen Kriterien und den ermittelten
individuellen Testwerten gibt und wenn ja wie groß dieser ist.
- Konstruktvalidität meint hingegen, dass die theoretisch fundierten Annahmen über
die Struktur und den Aufbau des Konstrukts formuliert und daraufhin empirisch
überprüft werden. Dadurch entsteht ein nomologisches Netzwerk auf die Frage „was
ein Test misst“.

Insgesamt handelt es sich bei der Validität um ein theoretisch komplexes Gebildet (Schmie-
mann & Lücken, 2014). Für die vorliegende Arbeit folgt daraus, dass es um Maßnahmen geht,
die für die Einschätzung der Validität des Tests herangezogen werden können. Entsprechend
werden Argumente hervorgebracht, die aufzeigen, „inwieweit die Interpretation der Ergeb-
nisse eines Tests gerechtfertigt sind“ (Hartig et al., 2008, S. 161).

6 ERGEBNISSE DER VORSTUDIE

6.1 KATEGORIENSYSTEM FÜR OFFENE AUFGABEN

Die Entwicklung und Validierung eines Kodiermanuals auf Basis eines Kategoriensystems zur Erfassung von Fähigkeiten im Umgang mit multiplen externen Repräsentationen (MER) im biologischen Kontext war Teil einer Masterarbeit (Schechert, unveröffentlicht, 2014). Aus diesem Grund werden in diesem Kapitel ausschließlich die zentralen Befunde des induktiv-deduktiv entwickelten Kategoriensystem vorgestellt und für genauere Hintergründe und Befunde auf die genannte Arbeit verwiesen.

Im Rahmen der deduktiven Kategorienbildung können die folgenden theoriebasierten Ergebnisse vergleichend zusammengefasst werden (vgl. TABELLE 9).

TABELLE 9: DEDUKTIVE KATEGORIENBILDUNG AUF BASIS DER THEORIEBASIERTEN INHALTSANALYSE

Kozma & Russell (2005)	Klöpfel et al. (2013)	Anderson et al. (2013)
Eigenschaften identifizieren (Fakt)	Innerhalb einer externen Repräsentation (ER) Elemente identifizieren (Fakt)	Elemente dekodieren (Fakt)
Eigenschaften beschreiben (Zusammenhang)		
Repräsentationen konstruieren (Zusammenhang)		Eigenschaften darstellen (Fakt, Zusammenhang)
Beziehung erklären (Zusammenhang, Konzept)	Innerhalb einer ER Zusammenhänge herstellen (Zusammenhang, zwei Zusammenhänge)	ER interpretieren und benutzen (Zusammenhang, zwei Zusammenhänge, Konzept)
Verbindung herstellen (zwei Zusammenhänge)	Informationen aus mehreren externen Repräsentationen (ERs) integrieren (Zusammenhang, zwei Zusammenhänge)	Horizontale Translation zwischen MERs eines Konzepts (Zusammenhang, zwei Zusammenhänge, Konzept)
Repräsentationen ineinander überführen (zwei Zusammenhänge)		
Konzepte und/ oder Phänomene beschreiben (Konzept)		

Vor dem Hintergrund der Einheitlichen Prüfungsanforderungen für das Abitur (EPA der KMK, 2004), werden zentrale Fähigkeiten den klassischen Anforderungsbereichen (I, II, III) zugeordnet und abschließend folgende Differenzierung als Grundlage der Kodierung festgehalten (vgl. TABELLE 10). Dabei wird einerseits nach dem Inhalt der Fähigkeit und andererseits nach der erwarteten Handlung mithilfe eines geeigneten Operators differenziert.

TABELLE 10: DEDUKTIVE FÄHIGKEITEN UND ANFORDERUNGSBEREICHE IM UMGANG MIT MER

Inhalt der Fähigkeit	Operator	identifizieren	beschreiben	erklären	interpretieren	begründen
Element/ Merkmal		AFB I	AFB I			
Eigenschaft		AFB I	AFB I			
Zusammenhang		AFB II	AFB II	AFB II	AFB II	AFB III
Konzept		AFB II	AFB II	AFB II	AFB II	AFB III

Anschließend folgte eine Erweiterung der bestehenden Kategorien durch das Einbeziehen von Kategorien, die im Rahmen des Strukturmodells zur Diagrammkompetenz (Lachmayer, 2008) validiert wurden. Die so gewonnen deduktiven Kategorien wurden als Ansatzpunkt für die induktive Kategorienbildung (vgl. Mayring, 2010) herangezogen und im Datenmaterial überprüft. Dadurch konnten weitere, induktiv gewonnene, Kategorien gefunden werden. Die Ergebnisse sind in TABELLE 11 dargestellt und dienen als Grundlage für die finale Erstellung des Kategoriensystems.

TABELLE 11: ERWEITERUNG ZU EINER INDUKTIV-DEDUKTIVEN KATEGORIENBILDUNG (MAYRING, 2010)

Informationsentnahme	Konstruktion
Wert markieren	Wahl des richtigen Diagrammtyps
Wert ablesen	Diagrammrahmen (DR) Achsenzuordnung
Charakteristisches Merkmal identifizieren (z.B. Maximum)	DR Achsenbeschriftung
Erkennen der dargestellten Relation	DR Achsenskalierung
Informationsentnahme (IE) aus dem Text	Kurve beschriften/ Legende
IE aus dem Schema	Werte eintragen
IE aus der Tabelle	Begriffe und/ oder Symbole eintragen
IE aus einer Kurve	Schema einzeichnen
IE aus mehreren Kurven	Trendlinie eintragen
Integration	
Zusammenhang erfassen	
Vergleich innerhalb einer ER und zwischen mehreren ERs	
Mathematisieren	
Zuordnen	
Abhängigkeiten erklären	
Kurvenverlauf in Verbindung zum Text beschreiben	
ER(s) zur Erklärung heranziehen	
Zusammenhang interpretieren/ begründen	

Im Anschluss daran folgte die finale Erstellung des Kategoriensystems sowie eines geeigneten Codierleitfadens für die Auswertung der Testaufgaben im offenen Antwortformat. Das Kategoriensystem geht dabei auf vier unterschiedliche Kompetenzausprägungen ein: Informationsentnahmenähe Integration (MERI), Konstruktionsnahe Integration (MERII), Integration aus mehr als zwei Repräsentationen (MERIII) als Translation und Transformation. Die Kategorien wurden auf die zwei unterschiedlichen Repräsentationstypen, Diagramm und Schemazeichnung, abgestimmt sowie an den bereits deduktiv und induktiv gewonnenen Einordnungen orientiert (vgl. TABELLE 10 und TABELLE 11). Jede Unterkategorie wurde definiert, beschrieben und mit einem Ankerbeispiel als Zitat aus dem Datenmaterial versehen. Das finale Kategoriensystem ist in TABELLE 12 dargestellt. Weitere Ausführungen finden sich in der vorgestellten Masterarbeit (vgl. Schechert, unveröffentlicht, 2014).

TABELLE 12: KATEGORIENSYSTEM ZUR CODIERUNG VON OFFENEN AUFGABEN MIT MER

Integrationsprozesse beim Umgang mit MER			
Kategorie	Informationsebnen- nahe Integration (MERI)	Konstruktionsnahe Integration (MERII)	Integration aus mehr als zwei Repräsentationen (Translation, MERIII)
Merkmale identifizieren = Meid	Markieren (ma) Wert ablesen (we) Charakteristisches Merkmal (cm) Informationsebnenahme aus mehreren Kurven (meku)	Informationsebnenahme (info) aus Text (te) oder Tabelle (ta) oder Schema (sch) Wert markieren (ma)	Informationsebnenahme (info) aus Text + zwei Diagramme Text + Diagramm + Schema Text + zwei Schemata Wert ablesen (we)
Merkmal(e)/ Eigenschaften beschreiben = Mebe	Kurven- oder Schemaverlauf (kv)	Wahl des richtigen Diagrammtyps (typ) Diagrammrahmen (dira): Variable zu Achsen zuordnen (dira_ac) Achsenbeschriftung (dira_bes) Zeichnen der Skalen (dira_ska) Kurve beschriften (ku) Legende (le) Punktwerte (we) Begriffe (be) oder Symbole (sy) einzeichnen/ ergänzen/ ausfüllen/ vollständigen	Diagrammrahmen (dira) Kurve beschriften (ku) Legende (le) Punktwerte (we) Begriffe (be) oder Symbole (sy) einzeichnen/ ergänzen/ ausfüllen/ vollständigen
Zusammenhang beschreiben = Zube	Zusammenfassen (zus) Vergleichen (ver) Mathematisieren (mat) Dargestellte Relation (rel) Zuordnen (zuo)	Trendlinie zeichnen (tre) Schema einzeichnen (sch) Überschrift/ Titel formulieren (übe) Mathematisieren (mat)	Trend einzeichnen (tre) Schema einzeichnen (sch) Überschrift formulieren (übe) Mathematisieren (mat)
		Kurven/ Säulen innerhalb einer ER vergleichen (ver) Kurven/ Säulen zwischen den MER vergleichen (ver_zw) Mathematisieren (mat) Zusammenfassen (zus) Dargestellte Relation (rel) Zuordnen (zuo)	

Zusammenhang erklären = Zuer	Abhängigkeiten erklären (er) Frage beantworteten (fra)	Erklären (er) Frage beantworteten (fra)	Informationen aus den MER erklären (er) Frage beantworteten (fra) Eine ER nutzen, um eine andere zu erklären (nu)	Erklären (er) Frage beantworteten (fra)
Zusammenhang interpretieren = Zuin	Interpretieren oder deuten (deu)		Interpretieren oder deuten (deu)	
Zusammenhang begründen = Zubeg	Begründen (beg)		Begründen (beg)	
Konzept beschreiben = Kobe		Ohne Information weiter, z.B. Extrapolieren oder Schema weiterzeichnen (wei)		Ohne Information weiter, z.B. Extrapolieren oder Schema weiterzeichnen (wei)
Konzept erklären = Koer	Generalisierten Zusammenhang erklären Frage beantworteten (fra)	Erklären (er) Frage beantworteten (fra)	Konzept mithilfe aller MER erklären (er_mer) Frage beantworteten (fra)	Erklären (er) Frage beantworteten (fra)
Konzept interpretieren = Koin	Interpretieren oder deuten (deu)		Informationen aus den MER unter einem Gesichtspunkt deuten (deu)	
Konzept begründen = Kobeg	Begründen (beg)		Begründen (beg)	

6.2 RASCH-SKALIERUNG: INTERNE KONSISTENZ DES KOMPETENZTESTS

Die Rasch-Skalierung wird zur Überprüfung der modellkonform konstruierten Testaufgaben herangezogen und im Folgenden für die Testung der zwei Itempools (A- und B-Serie) vorgestellt. Da die B-Serie für eine erweiterte komplementäre Itemgewinnung diente und für die Hauptstudie nur einzelne, ergänzende Items mit ausreichend guten Kennwerten aus dieser Serie ausgewählt wurden, wird die Ergebnisdokumentation der Rasch-Analyse für die insgesamt 80 untersuchten Testaufgaben nur im Überblick dargestellt.

Die 95 selektierten Testaufgaben der A-Serie (s. Kapitel 6.3) werden einer eindimensionalen, fallzentrierten Rasch-Skalierung unterzogen. Für die fallzentrierte Skalierung wird der Mittelwert der Personenfähigkeit auf Null gesetzt (s. Kapitel 5.2.2). In **ABBILDUNG 17** ist die Wright-Map dieser Skalierung für die 95 Testaufgaben vor der Itemselektion dargestellt.

Aufgrund der gemeinsamen Skalierung von Personen- und Itemparameter kann entschieden werden, ob der Personenparameter größer oder kleiner als der Itemparameter ist. Die Personen werden auf der Logit-Skala durch Kreuze (X) auf der linken Seite dargestellt. Rechts davon werden die Items, die durch Zahlen (1, 2, 3, ...) symbolisiert sind, aufgetragen. Im unteren Bereich der Wright-Map befinden sich Personen mit geringen Fähigkeitswerten sowie leichte Items. Personen mit hohen Kompetenzen sowie schwierige Items sind im oberen Bereich angesiedelt. Sind Person und Item auf einer Höhe, liegt die Lösungswahrscheinlichkeit der Person für dieses Item bei 50%, d.h. Studierende mit einem Personenwert von -1,0 Logits lösen beispielsweise das Item 45 mit einer 50-prozentigen Wahrscheinlichkeit. Items darüber werden mit geringerer und Items darunter mit höherer Wahrscheinlichkeit richtig gelöst (vgl. **ABBILDUNG 17**). Liegen Items über- oder unterhalb der angegebenen Personen (z.B. bei 2,0 Logits), deutet dies auf schlechte klassische Trennschärfen hin und die Items wären für die untersuchte Stichprobe deutlich zu schwer bzw. zu leicht. Als Folge davon werden solche Items aus dem Itempool entfernt.

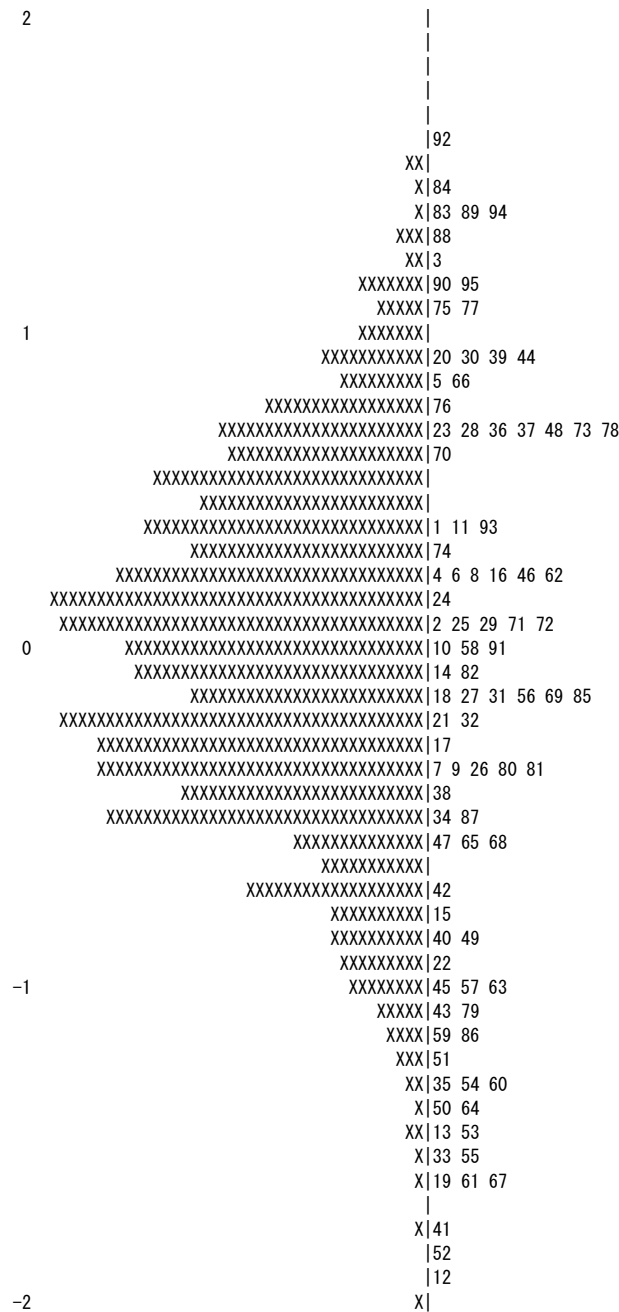


ABBILDUNG 17: EINDIMENSIONALE FALLZENTRIERTE RASCH-SKALIERUNG VOR SELEKTION (95 ITEMS DER A-SERIE; EIN X STEHT FÜR 0,6 STUDIERENDE)

In TABELLE 13 werden die zugehörigen Itemkennwerte der 95 Testaufgaben der A-Serie im Überblick dargestellt. Die empirisch ermittelten wMNSQ-Werte ($0,82 \leq wMNSQ \leq 1,15$) und T-Werte $-1,7 \leq T \leq 1,2$ liegen in den beschriebenen Grenzbereichen, während einige Items hinsichtlich der klassischen Trennschärfen keine ausreichende Passung zeigen.

TABELLE 13: ITEMKENNWERTE DER EINDIMENSIONALEN FALLZENTRIERTEN RASCH-SKALIERUNG VOR SELEKTION (95 ITEMS DER A-SERIE)

	Max.	Min.
wMNSQ	1,15	0,82
T-Werte	1,20	-1,70
r_{it}	0,74	-0,15

Aufgrund dieser Ergebnisse sowie im Rahmen der Kontextanalyse wurden 40 Testaufgaben aus dem Itempool der A-Serie entfernt (s. Kapitel 6.3). Für die verbliebenen 55 Testaufgaben liegen zwischen 21 und 59 Antworten je Item vor. Die Testaufgaben zeigen eine angemessene Schwierigkeit sowie zufriedenstellende Itemkennwerte. Mit diesen Items wurde erneut eine eindimensionale, fallzentrierte Rasch-Skalierung durchgeführt. Der Mittelwert der Personenfähigkeit wurde dabei auf Null gesetzt. Die Kennwerte der optimierten Skalierung nach Entfernung einiger Testaufgaben können TABELLE 14 entnommen werden. Die Wright-Map ist in ABBILDUNG 18 dargestellt.

TABELLE 14: ITEMKENNWERTE DER EINDIMENSIONALEN FALLZENTRIERTEN RASCH-SKALIERUNG NACH SELEKTION (55 ITEMS DER A-SERIE)

Skalierung			
Itemzahl	55		
Final Deviance	2678,25890		
Rel_i	0,88		
Varianz	0,80		
$Rel_{EAP/PV}$	0,43		
	S	Max.	Min.
wMNSQ		1,15	0,84
T-Werte		0,90	-1,10
r_{it}		0,73	0,25
I	3,81	1,80	-2,01
P	7,41	3,38	-4,03

Die empirisch ermittelten Kennwerte weisen auf eine gute Passung der Testitems auf das Rasch-Modell hin. Die wMNSQ-Werte liegen in einem Bereich von $0,84 \leq wMNSQ \leq 1,15$ und die T-Werte in einem Bereich von $-1,1 \leq T \leq 0,90$. Bei der klassischen Trennschärfe zeigen die verbliebenen Testaufgaben nach Selektion Werte von $r_{it} \geq 0,25$ auf.

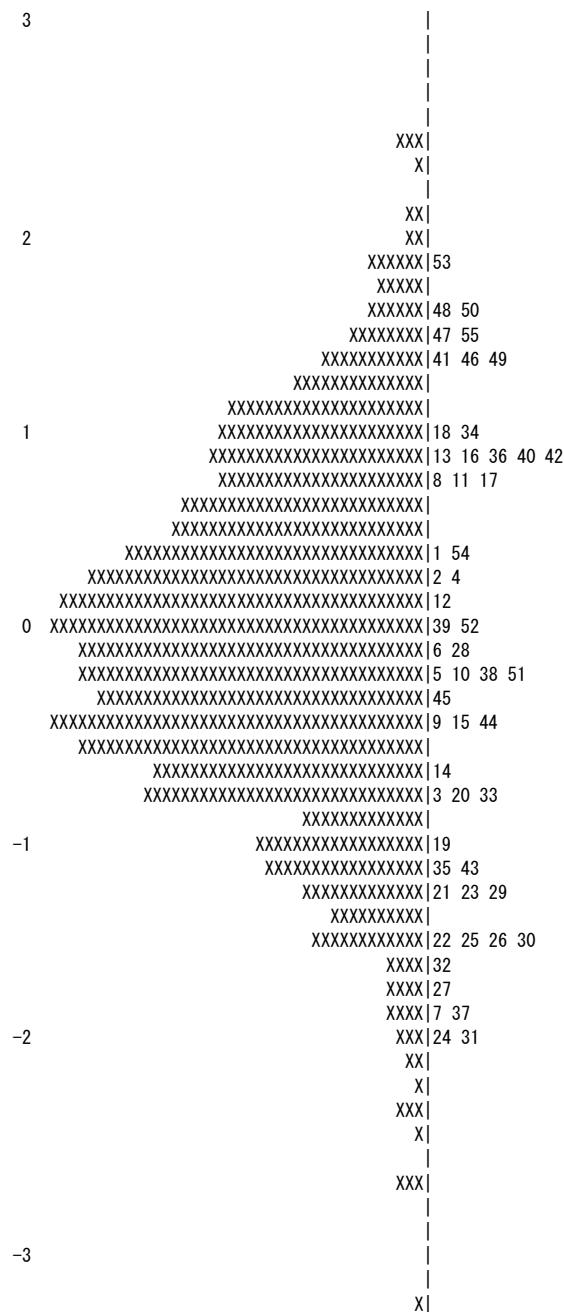


ABBILDUNG 18: EINDIMENSIONALE FALLZENTRIERTE RASCH-SKALIERUNG NACH SELEKTION (55 ITEMS DER A-SERIE, EIN X STEHT FÜR 0,6 STUDIERENDE)

Die ermittelten Personenfähigkeiten weisen eine Spanne von 6,18 Logits auf (Max.=3,38, Min.=-4,03). Die mittlere Personenfähigkeit beträgt 0,02 Logits ($SD=1,14$). Die Schwierigkeitsspanne der Itemparameter umfasst 3,81 Logits (Max.=1,80, Min.=-2,01), die mittlere Itemschwierigkeit liegt bei -0,13 Logits ($SD=0,46$). Die Spanne der Itemschwierigkeit ist insgesamt kleiner als die der Personenfähigkeiten. Besonders im oberen und unteren Bereich fehlen Items, die die Kompetenzen sehr leistungsstarker sowie sehr leistungsschwacher Studierender abdecken. Beim Vergleich der mittleren Personenfähigkeit und mittleren Itemschwierigkeit ist der Test insgesamt etwas zu leicht für die untersuchte Stichprobe. Die Item-Separation-Reliabilität Rel_i beträgt 0,88 und gibt die Zuverlässigkeit der Schätzung der

Itemparameter für die fallzentrierte Rasch-Skalierung an. Die *Rel_i* ist dabei von der Stichprobe abhängig und sollte sich mit zunehmender Größe dieser in der Hauptstudie weiter erhöhen. Die EAP/PV Person Separation Reliabilität gibt an, wie gut zwischen den Personen bezüglich ihrer Merkmalsausprägung unterschieden werden kann und steht im Zusammenhang mit der Aufgabengüte und der Anzahl von Testaufgaben. Die EAP/PV Reliabilität liegt mit 0,43 noch im moderaten Bereich, die Daten sollten dennoch vorsichtig interpretiert werden. Die Varianz ist ein Maß für die Streuung der Items um den Mittelwert und Voraussetzung für eine hohe Reliabilität (Brosius, 2013). Sie gibt damit die Breite der Kompetenzabdeckung an und liegt mit $\nu = 0,80$ in den vorliegenden Daten in einem mittleren bis hohen Bereich.

Abschließend wird die fallzentrierte eindimensionale Rasch-Skalierung der 90 Testaufgaben aus der B-Serie vorgestellt. Dabei erfolgt keine detaillierte Beschreibung der Itemselektion, da nur ausgewählte Items mit ausreichend guten Kennwerten in die Hauptstudie aufgenommen wurden. Unter Berücksichtigung des klassischen Schwierigkeitsindex lagen alle untersuchten Items der B-Serie in einem Schwierigkeitsbereich zwischen 20% und 80% Lösungswahrscheinlichkeit (Ausnahmen: Blu05, Ösy03, Cho06 als sehr schwere Items). Die empirisch ermittelten Kennwerte weisen auf eine gute Passung der Testitems auf das Rasch-Modell hin. Die *wMNSQ*-Werte liegen nach Selektion von drei Items (Dna02, Öko03 und See01) in einem Bereich von $0,80 \leq wMNSQ \leq 1,22$ und die *T*-Werte in einem Bereich von $-1,5 \leq T \leq 1,8$. Bei der klassischen Trennschärfe zeigen die Testaufgaben Werte $\geq 0,25$ auf, Items mit Werten $< 0,25$ wurden aus dem Itempool genommen (Pku04, Fot04, See10 und Cho02). Insgesamt liegt die klassische Trennschärfe in einem mittleren bis hohen Bereich mit Werten meist $\geq 0,40$. Die verbleibenden 73 Testaufgaben gingen in eine fallzentrierte eindimensionale Rasch-Skalierung ein. Die Wright-Map dieser Skalierung ist in ABBILDUNG 19 dargestellt. Die Item-Separation-Reliabilität (*Rel_i*) beträgt 0,64 und liegt wie die EAP/PV Person Separation Reliabilität mit einem Wert von 0,52 in einem moderaten Bereich. Die Varianz befindet sich mit $\nu = 0,66$ im mittleren Bereich. Vergleicht man die Itemparameter mit den Personenfähigkeiten in der Wright-Map selbst, ist der Kompetenztest der B-Serie insgesamt etwas zu schwer für die untersuchte Stichprobe.

Des Weiteren ist anzumerken, dass im unteren Kompetenz- und Itemschwierigkeitsbereich die Matrix lückenhaft ist und Items zur entsprechenden Kompetenzausprägung fehlen. Ebenfalls lassen sich hoch kompetente Studierende finden, für die Items fehlen, die ihre Kompetenz abdecken.

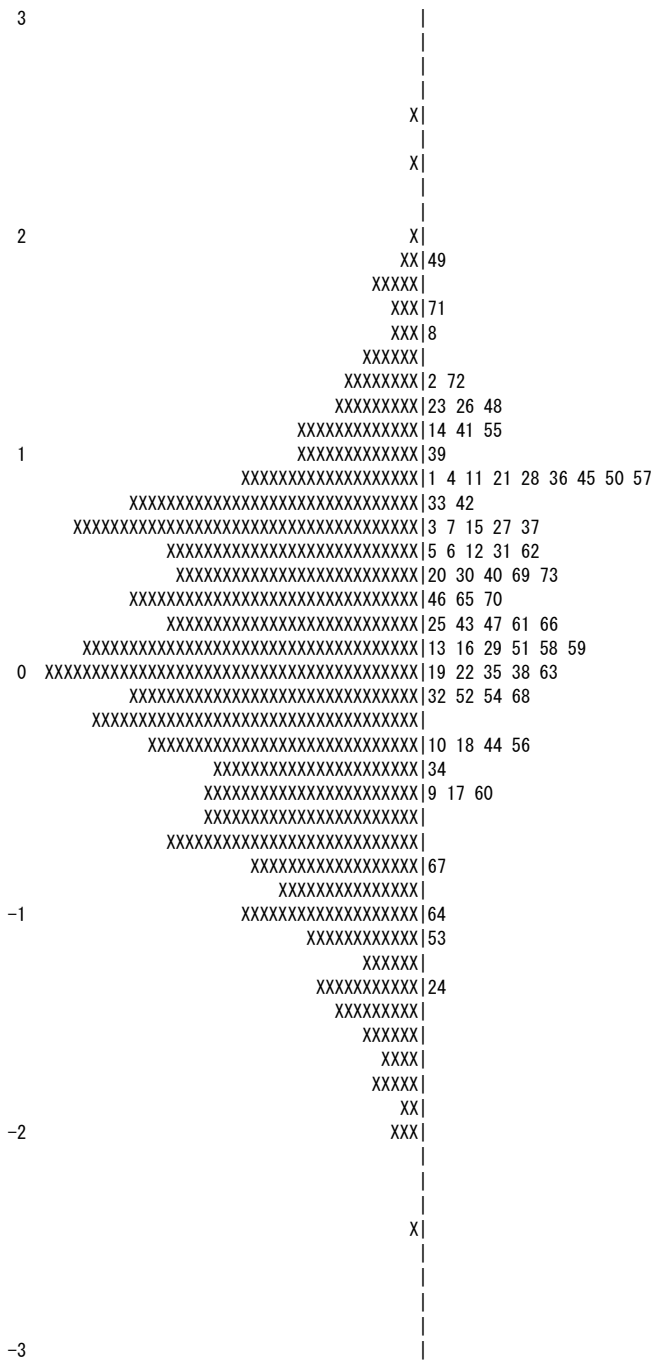


ABBILDUNG 19: EINDIMENSIONALE FALLZENTRIERTE RASCH-SKALIERUNG NACH SELEKTION (73 ITEMS DER B-SERIE, EIN X STEHT FÜR 0,5 STUDIERENDE)

Aufgrund der Testheftkonstruktion wird eine übergreifende Rasch-Analyse von A- und B-Serie vorerst nicht durchgeführt, sondern ausschließlich die Testung zweier Itempools präferiert (s. Kapitel 4.3). Eine genaue Beschreibung der Itemselektion erfolgt im nächsten Kapitel. Statistisch interessant sind erste Unterschiedsanalysen auf Itemebene, die in Kapitel 6.4 näher vorgestellt werden und sowohl die selektierten Items der A- als auch der B-Serie einbeziehen.

6.3 ITEMSELEKTION

Die A-Serie besteht insgesamt aus $n=105$ Testitems, wovon nach erster Sichtung bereits zehn Aufgaben aufgrund einer nicht eindeutigen Aufgabenstellung ausgeschlossen wurden und $n=95$ Testitems für die ersten Analysen resultierten (MERI-Items: Erb01_D, Kno01_D, Ösy01_S, Pku09_S; MERII-Items: Atm05_D, Hef06_D, Hef10_D, Kno08_D, Oly06_D, Wun01_S).

Anschließend wurden alle Testitems, die eine Antworthäufigkeit von weniger als 35 Antworten pro Item haben, genauer analysiert und direkt aus dem Itempool entfernt oder für eine weitere fachdidaktische Analyse markiert und ggf. später ausgeschlossen (MERI-Items: Atm10_S, Bir02_D, Blu10_D, Enz01_S, Enz03_D, Hef03_D, Reg03_D, Reg05_D; MERII-Items: Anp05_D, Bre05_D, Cho03_S; MERIII-Items: Blu12_S, Blu18_S, Dna04_S, Flo09_D, Flo10_D, Ösy07_S, Ösy10_S, See09_D). Im Anschluss daran folgte die deskriptive sowie probabilistische Analyse der verbleibenden Testaufgaben. Die Itemselektion erfolgte mit dem Ziel, am Ende möglichst wenig unterschiedliche Aufgabenstämme bei gleichzeitig möglichst großer Varianz in den Schwierigkeiten der Testitems zu haben. Damit konnte eine reduzierte Gesamtleszeit erreicht und der kognitiven Überlastung durch das Hineindenken in viele unterschiedliche Themenkomplexen entgegengewirkt werden. In einem ersten Schritt wurde zunächst der klassische Schwierigkeitsindex betrachtet. Alle Testitems mit einer Lösungswahrscheinlichkeit von mehr als 80 Prozent oder weniger als 20 Prozent wurden aus dem Itempool entfernt. Ausnahmen wurden nur dann zugelassen, wenn das Item einem etablierten Aufgabenstamm angehört, von besonderer fachdidaktischer Relevanz oder ein MERIII-Item war. Vor diesem Hintergrund wurden folgende Testitems ausgeschlossen oder zur weiteren Beobachtung markiert (MERI-Items: Bir05_D, Öko02_D; MERII-Items: Dna06_D, Öko04_D, Zuc03_D, Atm07_D; MERIII-Items: Ösy05_S).

In einem zweiten Schritt erfolgte die Analyse der Testitems mit einer Raschanalyse. Die weitere Itemselektion basierte anhand der wMNSQ-Werte, der T-Werte, der klassischen Trennschärfen (vgl. Bond & Fox, 2007) und anhand der Itemschwierigkeit. Daraufhin wurden folgende Items ausgeschlossen bzw. wiederholt markiert (vgl. TABELLE 15) und eine erneute Analyse mit 55 verbleibenden Testitems aus der A-Serie durchgeführt.

TABELLE 15: ITEMSELEKTION AUFGRUND DER RASCHANALYSE (N=55 ITEMS)

MER I	MER II	MER III
Ökologie: Flo01_D, Flo08_D Öko01_D, Öko02_D See02_D Zuc02_D Kno02_D, Kno03_D	Ökologie: Flo04_D Öko04_D See04_D, See07_D Zuc03_D, Zuc05_D	Ökologie: Flo07_D, Flo09_D, Flo10_D Öko06_D See09_D, Ösy08_D, Ösy10_S, Ösy05_S, Ösy07_S,
Stoffwechsel: Anp02_D, Anp04_D Atm04_D, Atm10_S Oly01_D, Oly03_S Enz01_S, Enz03_D	Stoffwechsel: Anp05_D Atm03_D Oly05_D, Oly08_D	Stoffwechsel: Anp07_S Atm01_S Blu02_S , Blu14_D Epo05_D
Genetik: Erb03_D Blu10_D Pku08_S	Genetik: Erb04_D Blu16_D Pku11_D	Genetik: Epo03_S Pku02_D, Pku06_S

Die B-Serie besteht insgesamt aus n=80 Testitems. Das Verfahren der Itemselektion erfolgte nach den gleichen Kriterien wie in Serie A berichtet. Aufgrund der bereits insgesamt 55 verbliebenden Testitems in der Serie A wurde die Analyse der B-Serie nur dahingehend verfolgt, dass einzelne Items zur Ergänzung des bestehenden Itempools identifiziert wurden. Nach erster Sichtung von klassischen und probabilistischen Kennwerten konnten ebenfalls Testitems ausgeschlossen bzw. zur weiteren Überprüfung markiert werden (Zuc04_D, Oly09_D, Ösy03_S, Blu05_S, Öko03_D, Dna02_S, Pku04_S, Fot04_D, See10_D, Cho02_S, Bir08_D, See01_S, Pku05_S, Ösy04_S, Rep02_D).

6.4 SCHWIERIGKEITSUNTERSCHIEDE

Im Folgenden werden die Ergebnisse der mehrfaktoriellen Varianzanalyse (ANOVA) vorgestellt, um zu untersuchen, ob sich die Mittelwerte der Itemschwierigkeiten der im Kompetenzstrukturmodell definierten Komponenten (MER-Integration, Aufgabenkontext, Repräsentationstyp) statistisch voneinander unterscheiden. Als unabhängige Variable werden jeweils die Kategorien „MER I, MER II oder MER III“ sowie „Ökologie, Stoffwechsel oder Genetik“ und „Diagramm oder Schema“ als Gruppierungsvariablen definiert und in Bezug auf die abhängige Variable „Lösungswahrscheinlichkeit“ untersucht. In ABBILDUNG 20 sind die mittleren Itemschwierigkeiten der fallzentrierten, eindimensionalen Rasch-Skalierung nach Komponenten der MER-Integration jeweils für die drei Aufgabenkontexte dargestellt. Als Diagramm

sind so genannte Boxplots dargestellt. Der dicke Strich in einem Balken entspricht dem 50%-Perzentil (Median).

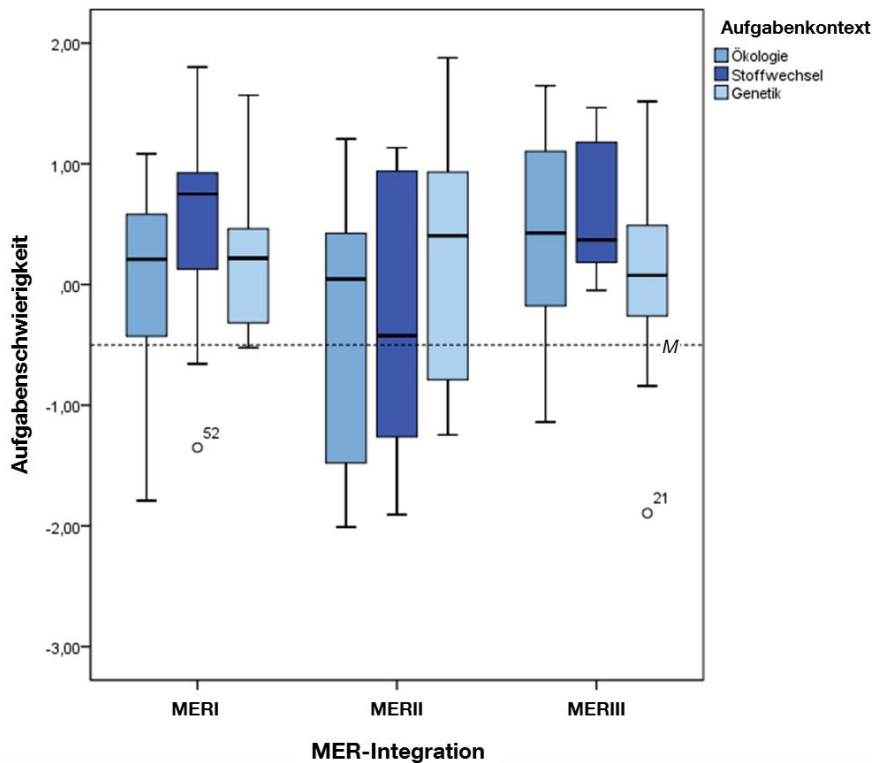


ABBILDUNG 20: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER MER-INTEGRATION FÜR DIE DREI AUFGABENKONTEXTE (N=128 ITEMS, M = MITTELWERT)

Für den Aufgabenkontext *Ökologie* sind Integrationsaufgaben der Komponente *MERIII* am schwierigsten zu lösen ($M_{MERIII}=0,44$, $SD=0,83$), gefolgt von *MERI*-Aufgaben ($M_{MERI}=0,01$, $SD=0,79$). Konstruktionsaufgaben waren in der Ökologie für die Stichprobe am leichtesten zu lösen ($M_{MERII}=-0,34$, $SD=1,10$). Für den Aufgabenkontext *Stoffwechsel* sind Integrationsaufgaben der Komponente *MERIII* ebenfalls am schwierigsten zu lösen ($M_{MERIII}=0,64$). Diese sind für die untersuchte Stichprobe auch im Kompetenztest insgesamt am schwierigsten. Für die Komponente *MERI* liegt die Schwierigkeit der Items bei einem Mittelwert von $M_{MERI}=0,60$ ($SD=0,65$). Konstruktionsaufgaben waren auch im Kontext *Stoffwechsel* für die Stichprobe am leichtesten zu lösen ($M_{MERII}=-0,39$, $SD=1,29$). Im Gegensatz dazu stehen die Aufgaben aus dem Aufgabenkontext *Genetik*. Mit einem Mittelwert von $M_{MERI}=0,24$ ($SD=0,63$) für die Integrationsaufgaben der Komponente *MERI* und einem Mittelwert von $M_{MERII}=0,18$ ($SD=1,06$) für die Konstruktionsaufgaben der Komponente *MERII* sind diese Aufgaben insgesamt leichter zu lösen. Integrationsaufgaben der Komponente *MERIII* werden von der untersuchten Stichprobe ebenfalls leicht gelöst ($M_{MERIII}=0,05$, $SD=0,98$).

Bezieht man in diese zweifaktorielle Analyse den Repräsentationstyp als dritten Faktor in das Modell ein, lassen sich zusammenfassend folgende Schwierigkeiten diagnostizieren (vgl. TABELLE 16). Damit können die Schwierigkeiten differenzierter interpretiert werden.

TABELLE 16: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER MER-INTEGRATION, GETRENNT NACH AUFGABENKONTEXTEN UND DEM REPRÄSENTATIONSTYP (N=128 ITEMS)

		MER I		MER II		MER III	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Ökologie	Diagramm	-0,04	0,83	-0,36	1,10	0,17	0,76
	Schema	0,40	0,26	-0,06	1,00	0,98	0,73
Stoffwechsel	Diagramm	0,61	0,60	-0,42	1,22	0,77	0,74
	Schema	0,60	0,77	-0,32	1,05	0,55	0,62
Genetik	Diagramm	0,28	0,74	-0,22	0,93	-0,05	1,23
	Schema	0,17	0,53	0,77	1,08	0,15	0,81

Der differenzierte Blick auf die Itemschwierigkeiten zeigt, dass neben der MER-Integration und dem Aufgabenkontext auch der Repräsentationstyp eine Rolle im Hinblick auf die Itemschwierigkeit spielt. So sind es vor allem Diagrammaufgaben in der Stoffwechselphysiologie, die die Informationsentnahmenähe Integration schwierig machen ($M_{MERISD}=0,61$, $SD=0,60$). Gleichzeitig ist die konstruktionsnahe Integration von Diagrammaufgaben in der Stoffwechselphysiologie am leichtesten ($M_{MERIISD}=-0,42$, $SD=1,22$). Des Weiteren kann für die MERII-Integration festgehalten werden, dass die Testaufgaben insgesamt am leichtesten für die untersuchte Stichprobe zu lösen waren, eine Ausnahme stellen MERII-Aufgaben im Kontext Genetik dar, die eine Integration mit dem Repräsentationstyp Schema erfordern ($M_{MERIIS}=0,77$, $SD=1,08$). Während MERIII-Aufgaben aus der Stoffwechselphysiologie insgesamt am schwierigsten zu lösen waren, zeigt der Blick auf den Repräsentationstyp, dass Schemaaufgaben im Kontext Ökologie noch schwieriger zu lösen waren, als die Diagrammaufgaben im Kontext Stoffwechsel ($M_{MERIIIÖS}=0,98$, $SD=0,73$). Die Ergebnisse implizieren, dass eine zusätzliche qualitative Untersuchung der Testaufgaben einen Beitrag zur Interpretation der MER-Integrationsprozesse leisten kann. In diesem Zusammenhang sei auf die qualitative Analyse der Testaufgaben im Rahmen der Hauptstudie verwiesen (s. Kapitel 9.4).

Über die mehrfaktoriellen Varianzanalyse (ANOVA) wird überprüft, ob sich die Mittelwerte der Itemschwierigkeiten aus dem 3 x 3 x 2-Design des Kompetenzstrukturmodells auch statistisch signifikant voneinander unterscheiden. Normalverteilung und Varianzhomogenität als Voraussetzung für die Durchführung der Varianzanalyse wurden kontrolliert, der Levene-Test ist nicht signifikant ($F(17,110) = 1,551, p = 0,09$). Der Test ist insgesamt signifikant, d.h. das Gesamtmodell ist signifikant ($F(17,110) = 1,737, p = 0,046$, partielles $\eta^2=0,21$). Als Anteil der Streuung um den Gesamtmittelwert können 9 % Varianz durch das Modell erklärt werden (korr. R-Quadrat = 0,09). Im Hinblick auf die unterschiedlichen Komponenten gibt es im mehrfaktoriellen Modell hingegen keinen Haupteffekt der MER-Integration, des Aufgabenkontextes oder des Repräsentationstyps auf die Itemschwierigkeit ($p > 0,05$). Aus diesem Grund wurden zusätzlich einfaktorielle Varianzanalysen durchgeführt, indem jede Komponente aus dem Kompetenzstrukturmodell als eine unabhängige Variable definiert wurde (z.B. MER-Integration) und die Schwierigkeitsunterschiede in dieser Gruppe analysiert wird. Da der Levene-Test signifikant ist ($F(2,125) = 6,645, p = 0,002$) und damit keine Varianzhomogenität vorliegt, wurde der Welch-Test herangezogen. Als Ergebnis resultiert, dass der Test signifikant ist ($F(2,125) = 5,656, p = 0,004$) und ein Haupteffekt der MER-Integration auf die Itemschwierigkeit vorliegt. Multiple Mittelwertsvergleiche im Rahmen der Post-Hoc-Analyse mit Bonferroni-Korrektur zeigen, dass signifikante Unterschiede zwischen den Komponenten MERI und MERII (0,57, 95%-CI [0,11, 1,02]; $p < 0,05$) und zwischen den Komponenten MERII und MERIII (-0,59, 95%-CI [-1,08, -0,11]; $p < 0,05$) vorliegen. Die einfaktorielle Varianzanalyse der Aufgabenkontexte ergab keine statistischen Unterschiede in den Itemschwierigkeiten von Ökologie, Stoffwechsel oder Genetik ($F(2,125) = 0,862, p = 0,425$). Zur Berechnung der Schwierigkeitsunterschiede im Repräsentationstyp wurde ein t-Test für unabhängige Stichproben durchgeführt. Analog zur Varianzanalyse setzt der t-Test Varianzhomogenität aus. Diese wurde überprüft, der Levene-Test ist nicht signifikant ($F(1,126) = 1,418, p = 0,236$). Die Teststatistik beträgt $t = -2,360$ (-0,40, 95%-CI [-0,74, 0,06]; $p = 0,02$) bei einer Effektstärke von $r = 0,21$. Damit liegt ein signifikanter Unterschied zwischen dem Repräsentationstyp Diagramm und Schema vor.

Inwieweit der Repräsentationstyp innerhalb der verschiedenen MER-Integrationsleistungen sowie innerhalb den verschiedenen Aufgabenkontexten eine Rolle spielt, wurde in Ergänzung ebenfalls untersucht. Die Ergebnisse der Schwierigkeitsverteilungen sind in ABBILDUNG 21 und ABBILDUNG 22 dargestellt.

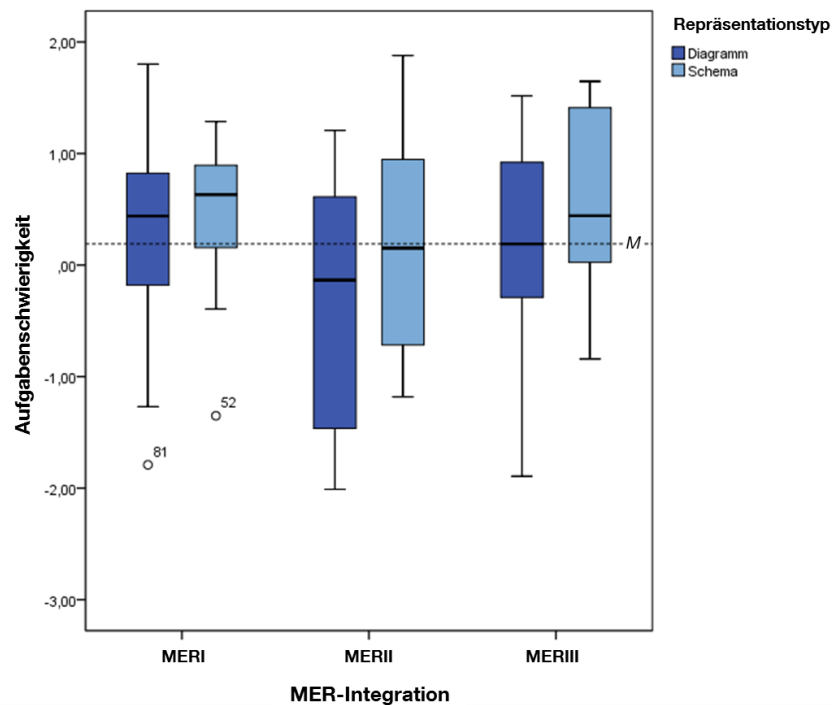


ABBILDUNG 21: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER MER-INTEGRATION FÜR DIE ZWEI REPRÄSENTATIONSTYPEN (N=128 ITEMS, M = MITTELWERT)

Das Schwierigkeitsprofil zeigt insgesamt, dass Schemaaufgaben tendenziell über alle MER-Integrationsprozesse schwieriger zu lösen sind als Diagrammaufgaben. Am leichtesten sind Testaufgaben zur *MERII-Integration*, die Diagramme enthalten ($M_{MERIID}=-0,35$, $SD=1,07$). Auch die Schemaaufgaben der *MERII-Integration* sind mit einem Mittelwert von $M_{MERIIS}=0,19$ ($SD=1,07$) leichter als die anderen Items der Komponenten MERI oder MERIII. Testaufgaben mit Diagrammen, die eine *Informationsentnahmenähe Integration* erfordern ($M_{MERID}=0,29$, $SD=0,76$), sind hingegen schwieriger als MERIII-Aufgaben ($M_{MERIID}=0,20$, $SD=0,87$). Am schwierigsten zu lösen sind Testaufgaben der *MERIII-Integration*, die ein Schema enthalten ($M_{MERIIIS}=0,61$, $SD=0,78$), gefolgt von MERI-Aufgaben ebenfalls mit einer Schemazeichnung ($M_{MERIS}=0,47$, $SD=0,67$). Die anschließende zweifaktorielle Varianzanalyse ist statistisch signifikant ($F(5,122) = 3,330$, $p = 0,007$, partielles $\eta^2=0,12$), zusätzlich liegt ein Haupteffekt des Repräsentationstyps auf die Itemschwierigkeit vor ($p < 0,05$).

Hinsichtlich der unterschiedlichen Aufgabenkontexte zeigt sich das folgende Bild der Verteilung der Itemschwierigkeiten (vgl. ABBILDUNG 22).

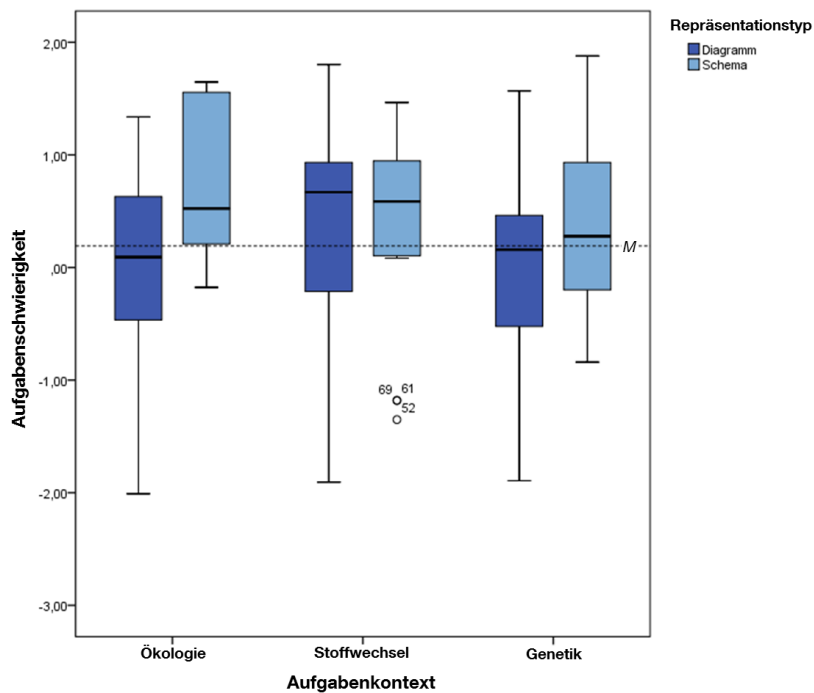


ABBILDUNG 22: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER AUFGABENKONTEXTE FÜR DIE ZWEI REPRÄSENTATIONSTYPEN DIAGRAMM UND SCHEMA (N=128 ITEMS, M = MITTELWERT)

Schemaaufgaben innerhalb des Kontextes Ökologie mit einem Mittelwert von $M_{\text{ÖkoS}}=0,76$ ($SD=0,23$) sind schwieriger für die untersuchte Stichprobe zu lösen als Testaufgaben mit Diagrammen ($M_{\text{ÖkoD}}=-0,07$, $SD=0,14$). Im Kontext Stoffwechsel sind Diagrammaufgaben schwieriger zu lösen ($M_{\text{StwD}}=0,29$, $SD=0,18$) als im Kontext Ökologie und im Kontext Genetik, aber leichter als Schemaaufgaben mit einem Mittelwert von $M_{\text{StwS}}=0,38$ ($SD=0,20$). Ein ähnliches Bild zeigt sich im Kontext Genetik. Schemaaufgaben sind mit einem Mittelwert von $M_{\text{GenS}}=0,35$ ($SD=0,23$) schwieriger zu lösen als Diagrammaufgaben ($M_{\text{GenD}}=0,01$, $SD=0,23$). Ergebnisse einer zweifaktoriellen Varianzanalyse mit den Gruppierungsvariablen Repräsentationstyp und Aufgabenkontext zeigen nach Überprüfung der Eingangsvoraussetzungen (Normalverteilung, Varianzhomogenität) hingegen keine statistisch signifikanten Effekte. Die Items 52, 60 und 61 stellen Ausreißer im Boxplot-Diagramm dar, werden aber aufgrund der Überprüfung der Testaufgaben als Items für die Hauptstudie für diese Darstellung der Ergebnisse nicht entfernt.

Aufgrund der sichtbaren Schwierigkeitsunterschiede, den allerdings nicht auch statistisch vorliegenden signifikanten Unterschieden in den Varianz- und Posthoc-Analysen, ist eine qualitative Analyse der Testaufgaben anzustreben. Diese kann ein detaillierteres Bild der Schwierigkeitsverteilung im Hinblick auf Anforderung, Aufgabenstellung und MER-Integrationsleistung in den unterschiedlichen Aufgabenkontexten liefern und ist Bestandteil der Ergebnispräsentation im Rahmen der Hauptstudie (s. Kapitel 9.4).

6.5 ZUSAMMENFASSUNG UND IMPLIKATIONEN FÜR DIE HAUPTSTUDIE

Für die vorliegende Untersuchung wurde ein Itempool von $n=210$ Aufgaben unterschiedlicher Antwortformate geschaffen (s. Kapitel 4.2.2), wovon 186 Testaufgaben in die Pilotierung ein gingen. Die hohe Anzahl offener Aufgaben führte zur Entwicklung und Validierung eines Kategoriensystems zum Umgang mit MER (s. Kapitel 6.1). Im Anschluss an die Kodierung im „Full-/ No-Credit-System“ erfolgte eine Analyse im Rahmen von deskriptiven Statistiken und der Rasch-Skalierung (s. Kapitel 6.2). Dieses Vorgehen diente der Evaluation der entwickelten Testitems auf ihre Schwierigkeit, die statistische Güte und ihre Verwertbarkeit als Testaufgaben für die anschließende Durchführung der Hauptstudie.

In der Pilotierung liegen für die A-Serie durchschnittlich $n=42,7$ Antworten, für die B-Serie durchschnittlich $n=37,0$ Antworten pro Testaufgabe vor. Aufgrund von weniger als 100 Antworten (vgl. Senkbeil et al., 2005) pro Testitem wurden keine Analysen auf Personenebene, sondern ausschließlich Analysen der eindimensionalen Raschanalyse durchgeführt. Die statistischen Kennwerte zeigen, dass die modellkonform konstruierten Testaufgaben den Qualitätskriterien genügen und zur Abbildung der Struktur des postulierten Kompetenzmodells zum Umgang mit multiplen externen Repräsentationen (MER) geeignet sind. Die hohen Trennschärfen decken sich mit der in der Literatur genannten Annahme, dass offene Antwortformate messtheoretisch eine höhere Trennschärfe zeigen im Gegensatz zu Multiple-Choice-Items (u.a. Klieme et al., 2000; Grube, 2010).

Für eine genauere Schätzung der Itemparameter in der Hauptstudie ist auf eine Erhöhung der Antworthäufigkeit pro Item zu achten (vgl. Linacre, 1994). Das Fähigkeitsspektrum der Studierenden konnte insgesamt nicht vollständig mit dem Itempool abgedeckt werden, da Items im unteren und oberen Schwierigkeitsbereich fehlen. Die EAP/PV-Reliabilität ist moderat, aufgrund der geringen individuellen Datengrundlage musste mit Reliabilitätseinbußen gerechnet werden (Senkbeil et al., 2005). Eine Reliabilitätsverbesserung des Kompetenztests wäre für die Hauptstudie hingegen wünschenswert. Die Varianzen der Skalen sind zufriedenstellend. Aus diesem Grund ist es sinnvoll zusätzlich methodenspezifische Einflussfaktoren wie das Testheftdesign sowie die Testaufgaben im offenen Antwortformat als Hintergrundvariablen zur Beurteilung der Reliabilität zu berücksichtigen. Verschiedene Studien (u.a. Klieme et al., 2000; Leucht et al. (2012); Härtig et al., 2015) führen an, dass offene Antwortformate zu niedrigeren Personenfähigkeiten führen als Multiple-Choice-Aufgaben. Stecher & Klein (1997) sowie die Ergebnisse der PISA-Studien 2003 und 2006 (vgl. Kulgemeyer & Schecker, 2007; OECD, 2009) betonen, dass die Reliabilitäten mit zunehmendem Anteil offener

Antwortformate sinken. Leucht et al. (2012) berichten, dass der Einfluss des offenen Aufgabenformats häufig vergleichsweise hoch ist, zur Validierung von Kompetenzstrukturen jedoch ein möglichst geringer Einfluss oder dieser konstant gehalten werden sollte. Im Rahmen der Vorstudie wurden jeder Testperson 18 Items zur Beantwortung vorgelegt. Die beschriebene Anwohthäufigkeit pro Item zeigt hingegen, dass die maximale Anzahl der Testaufgaben pro Testheft reduziert werden muss. Die Testzeit sollte gekürzt werden, damit ist eine adäquate Itemauswahl mit möglichst wenigen Aufgabenstämmen notwendig, die dennoch eine Abbildung des Kompetenzstrukturmodells über die verschiedenen Komponenten garantiert. An dieser Stelle sei angemerkt, dass die Itemzahl pro Testheft auf neun verringert wurde (s. Kapitel 7.2). Dadurch sollte eine immer noch zufriedene und genaue Schätzung der Personenparameter möglich sein (vgl. Rost, 2004; s. Kapitel 5.2.2).

Hinsichtlich der Itemschwierigkeiten zeigen sich über die untersuchten Komponenten deutliche Unterschiede. Dabei sind neben der *Integration von MER* auch der *Aufgabenkontext* und der *Repräsentationstyp* entscheidend. Für das Gesamtmodell, bestehend aus Integration von MER, Aufgabenkontext und Repräsentationstyp, waren Diagrammaufgaben vor allem im Kontext Stoffwechsel bei der Informationsentnahmenahen Integration (MERI) herausfordernd. Testaufgaben, die eine konstruktionsnahe Integration erfordern, waren vor allem in der Genetik schwierig, wenn Schemaaufgaben eingesetzt wurden. Für den Kontext Ökologie waren ebenfalls Schemaaufgaben am schwierigsten, vor allem wenn eine Integration aus dem Bereich MERIII gefordert war. Die Schwierigkeitsverteilung ändert sich hingegen, wenn unterschiedliche Komponenten einzeln in den Blick genommen werden. Aus diesem Grund sollten die beiden Komponenten *Aufgabenkontext* und *Repräsentationstyp* auch im Rahmen der Hauptstudie berücksichtigt und differenzierter analysiert werden. Dabei wäre es wünschenswert die aufgeklärte Varianz im Gesamtmodell durch die einzelnen Komponenten des Kompetenzstrukturmodells insgesamt zu erhöhen. Aufgrund der statistischen Kennwerte der Rasch-Skalierung und den mittleren Itemschwierigkeiten wurden einige der konstruierten Testaufgaben aus dem Itempool entfernt (s. Kapitel 6.3). Vor dem Hintergrund der resultierenden Ergebnisse sowie den methodischen Grenzen und der inhaltlich sehr breiten Auswahl verschiedener Aufgabenkontexte, muss die Itemzahl für die Hauptstudie deutlich reduziert werden. Entscheidend dabei ist, dass unter Berücksichtigung der drei Teilkomponenten (MERI, MERII, MERIII) eine adäquate Auswahl von Aufgabenkontexten erfolgt, die zusätzlich beide Repräsentationstypen berücksichtigt (s. Kapitel 7.2).

Damit lassen sich aus den Vorstudienresultaten folgende Implikationen für die Durchführung und Analyse der Hauptstudie ableiten:

- I. Optimierung der psychometrischen Qualität des Kompetenztests,
- II. Dimensionalitätsprüfungen der unterschiedlichen mehrdimensionalen Modelle im Vergleich zum eindimensionalen Rasch-Modell,
- III. Prüfung des Einflusses der Dimensionen (MER-Integration) und der Komponenten Aufgabenkontext und Repräsentationstyp auf die Aufgabenschwierigkeit,
- IV. Untersuchung der Kompetenzunterschiede und Zusammenhänge erfasster Kompetenzen mit weiteren Variablen.

7 MATERIAL DER HAUPTSTUDIE

7.1 STICHPROBE UND ORGANISATION DER TESTUNG

Die Population der Hauptstudie setzte sich aus $N=548$ Studierenden des zweiten Semesters und unterschiedlicher Fachrichtungen zusammen (♀ 48,2 %, ♂ 51,8 %; Alter $M=20,66$, $SD=2,55$). Ablauf und Organisation fanden analog zur Vorstudie statt (s. Kapitel 4.4). An der Untersuchung haben folgende Universitäten teilgenommen: Technische Universität München (TUM), Ludwig-Maximilians-Universität München (LMU), die Universität zu Köln sowie die Christian-Albrechts-Universität zu Kiel (CAU). Die Verteilung der Studierenden und ihre Studiengänge, die im Rahmen der Hauptstudie beteiligt waren, können TABELLE 17 entnommen werden.

TABELLE 17: HAUPTSTUDIE - TEILGENOMMENE STUDIENGÄNGE UND ANZAHL DER STUDIERENDEN (N=548)

Studiengang	Anzahl der Studierenden
Lehramt Gymnasium/ Realschule/ Mittelschule mit Fach Biologie	83
Lehramt Gymnasium/ Realschule/ Mittelschule mit Fach Mathematik	25
Lehramt Gymnasium/ Realschule/ Mittelschule mit Fächerkombination Biologie & Mathematik oder Biologie & Sport	9
Biowissenschaften	137
Sonstige Natur- und Ingenieurwissenschaften	169
Mathematik (inkl. Physik und Informatik)	98
Sonstige oder keine Angaben	9

Um die Geltung eines Modells zu testen, sollte eine entsprechende Stichprobengröße vorliegen. Rost (2004) schlägt einen Stichprobenumfang von ca. 100 Probanden vor, Wright & Stone (1999) schlagen mindestens 200 Probanden vor, wenn die zu bearbeitenden Testhefte über zehn gute Ankeritems verlinkt sind. Die Hauptstudie beinhaltete $n=36$ Testaufgaben, von denen jeder Testperson neun Items zur Beantwortung vorgelegt wurde. Jedes Item wurde bei der vorliegenden Stichprobe von $N=548$ durchschnittlich von 95-131 Studierenden beantwortet. Damit kann eine zufriedenstellende Schätzung der Item- und Personenkennwerte angenommen werden.

7.2 TESTKONSTRUKTION

In der Hauptstudie geht es darum, die Modellannahmen zu überprüfen. Aus diesem Grund müssen diejenigen Testitems ausgewählt werden, die geeignet sind, das zu untersuchende Konstrukt abzubilden. Ziel der Itemzusammenstellung ist damit ein ausreichend großer Itempool an hochwertigen Testaufgaben, die die gesuchte Kompetenz valide abbildet sowie reliable und objektive Messergebnisse produziert (vgl. Downing & Haladyna, 1997). Aus den bereits diskutierten Gründen (s. Kapitel 6) wurden für die Durchführung der Hauptstudie 36 Items festgelegt, die sich im Rahmen der Pilotierung als statistisch und inhaltlich geeignet erwiesen haben.

Aufgabenkonstruktion: In der Hauptstudie wurde die Beantwortung der ausgewählten Testaufgaben mit einer maximalen Bearbeitungszeit von 45 Minuten pro Testheft realisiert. Daraufhin mussten die pilotierten Testitems gekürzt werden. Unter Beibehaltung der charakteristischen Elemente in der Itembeschreibung (s. Kapitel 4.2.2) wurde das Glossar zu den verwendeten Fachbegriffen entfernt und der Aufgabenstamm in seiner Länge reduziert. In diesem Zusammenhang wurde berücksichtigt, dass möglichst wenig verschiedene Aufgabenstämme enthalten sind, um die Lesezeit nicht unnötig zu erhöhen. Aufgabenstimulus und Problemaufriss blieben erhalten, damit der Testperson eine adäquate Lösung der Aufgabe vorgegeben werden kann (vgl. Hartig & Jude, 2007). Auf eine präzise Formulierung der Fragestellung unter Verwendung passender Operatoren wurde zusätzlich geachtet. Bei jeder Itemanalyse stand zudem die fachdidaktische Perspektive und psychometrisch adäquate Umsetzung davon im Vordergrund. Wie in der Vorstudie wurden auch Alter, Studiengang, das Geschlecht sowie die Biologie- und Mathematiknote als Selbstauskunft erhoben (s. Kapitel 4.2).

Itemumfang: Für die Verteilung der Items auf die Testhefte wurde aufgrund der positiven Erfahrungen sowie aus testökonomischen Gründen erneut das Youden-Square-Design angewendet (s. Kapitel 7.3). In diesem Zusammenhang wurden die verbleibenden Testitems analog zur Vorstudie zu Clustern aggregiert. Um allen Probanden eine realistische Möglichkeit der Beantwortung zu geben, wurden pro Testheft neun Items verteilt. Für jede Teilkomponente (MERI, MERII, MERIII) und Facette im Kompetenzmodell (Aufgabenkontext, Repräsentationstyp) wurden drei adäquate, valide offene Testaufgaben ausgewählt. Damit diese Umsetzung gelingt, wurde sich dafür entschieden, die Verteilung der Testitems innerhalb der Facette Repräsentationstyp zu reduzieren. Daraus folgt, dass für den Aufgabenkontext Ökologie ausschließlich Testaufgaben mit Diagrammen verwendet wurden (9 Items). Für den Auf-

gabenkontext Genetik wurden Schemazeichnungen mit Stammbäumen und fachspezifischen konventionalisierten Repräsentationen ausgewählt (9 Items). Um den Vergleich innerhalb eines Aufgabekontextes zu erhalten, wurden im Kontext Stoffwechsel beide Repräsentationstypen aufgenommen, d.h. es waren Aufgaben mit Diagrammen als auch Aufgaben mit Schemazeichnungen vorgesehen (18 Items). Dadurch konnte die Itemzahl pro Aufgabekontext reduziert werden, gleichzeitig erlaubt die Itemverteilung von n=12 Testitems je Teilkomponente der MER-Integration eine messtechnisch zufriedenstellende Voraussetzung (vgl. TABELLE 18).

TABELLE 18: HAUPTSTUDIE- VERTEILUNG DER TESTAUFGABEN (ITEMS) FÜR DAS KOMPETENZKONSTRUKT "UMGANG MIT MER" IN DREI AUFGABENKONTEXTEN UND ZWEI REPRÄSENTATIONSTYPEN (N=36 ITEMS)

Umgang mit multiplen externen Repräsentationen																	
MERI						MERII						MERIII					
Öko		Stw		Gen		Öko		Stw		Gen		Öko		Stw		Gen	
3 Items		6 Items		3 Items		3 Items		6 Items		3 Items		3 Items		6 Items		3 Items	
D	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S
3	/	3	3	/	3	3	/	3	3	/	3	3	/	3	3	/	3
12 Testaufgaben						12 Testaufgaben						12 Testaufgaben					

Für das in Kapitel 7.3 herangezogene Testheftdesign wurden die 36 Testitems geclustert und so auf die Teilkompetenzen aufgeteilt, dass eine ausgeglichene Schwierigkeitsverteilung vorliegt. Zusätzlich wurde sichergestellt, dass in jedem Testheft mindestens ein Cluster, bestehend aus drei Items, aus dem Aufgabekontext Stoffwechsel stammt. Des Weiteren wurde darauf geachtet, dass in jedem Testheft alle drei Teilkompetenzen (MERI, MERII, MERIII) in allen drei Aufgabekontext repräsentiert sind. Die Verteilung kann TABELLE 19 entnommen werden. Vor diesem Hintergrund wird deutlich, dass sich nur eine ganz bestimmte Auswahl der pilotierten Testitems für eine derartige Itemverteilung als geeignet erwies.

TABELLE 19: HAUPTSTUDIE - VERTEILUNG DER ITEMS AUF DIE MER-TEILKOMPONENTEN (I, II, III) UNTER BERÜCKSICHTIGUNG VON AUFGABENKONTEXT UND REPRÄSENTATIONSTYP

Cluster	Item 1	Item 2	Item 3	Cluster	Item 1	Item 2	Item 3	Typ
1	Öko I	Gen II	Öko III	2	SW I	SW II	SW III	D
3	Gen I	Öko II	Gen III	4	SW I	SW II	SW III	S
5	Gen III	Öko I	Gen II	6	SW III	SW I	SW II	D
7	Öko III	Gen I	Öko II	8	SW III	SW I	SW II	S
9	Öko II	Gen III	Öko I	10	SW II	SW III	SW I	D
11	Gen II	Öko III	Gen I	12	SW II	SW III	SW I	S

Zusammenfassend wird damit die Auswahl folgender Testaufgaben aus dem konstruierten Itempool für den Einsatz in der Hauptstudie festgehalten (vgl. TABELLE 20).

TABELLE 20: HAUPTSTUDIE – FINALE ITEMSAUSWAHL, GETRENNT NACH MER-INTEGRATION, AUFGABENKONTEXT UND REPRÄSENTATIONSTYP

Umgang mit multiplen externen Repräsentationen

MERI			MERII			MERIII		
Öko	Stw	Gen	Öko	Stw	Gen	Öko	Stw	Gen
D	D & S	S	D	D & S	S	D	D & S	S
See02	Anp03	Blu04	See07	Atm03	Epo03	See09	Apr01	Erb05
Öko02	Anp04	Blu09	Öko04	Epo03	Blu05	Öko06	Blu11	Blu08
Flo08	Atm04	Dna02	Flo04	Epo05	Pku08	Flo09	Blu15	Pku07
	Atm11			Blu06			Blu02	
	Anp09			Enz05			Anp07	
	Enz04			Enz07			Enz02	

Antwortformat: Vor diesem Hintergrund änderte sich die Verteilung der Antwortformate in der Hauptstudie. Während die Multiple-Choice-Items fast vollständig ausgeschlossen wurden, besaß etwa zwei Drittel der Testitems ein offenes ausführliches Antwortformat, während mehr als ein Drittel der Testitems auf Items mit offenem Kurzantwortformat fiel (vgl. TABELLE 21). Die Kodierung der Antworten erfolgte analog zur Vorstudie im „Full-/ No-Credit-System“ (s. Kapitel 4.2.2).

TABELLE 21: HAUPTSTUDIE - ITEMS IN UNTERSCHIEDLICHEN ANTWORTFORMATEN

	Hauptstudie							
	MERI		MERII		MERIII		Gesamt	
	N	%	N	%	N	%	N	%
geschlossen – Multiple Choice	1	2,77	/	0	1	2,77	2	5,55
offen – Kurzantwort (Kurzantwort-, Ablese-, Zuordnungsaufgaben)	2	5,55	7	19,44	4	11,11	13	36,11
offen – ausführliche Antwort (Operatoren)	9	25,00	5	13,88	7	19,44	21	58,33
Gesamt	12	33,33	12	33,33	12	33,33	36	100

7.3 TESTHEFTDESIGN

Im Rahmen der Hauptstudie wurde sich erneut am Youden-Square-Design (YSD) und am Design lateinischer Quadrate orientiert. Daraufhin wurde ein angepasstes Design gewählt, um analog zur Vorstudie die Positionseffekte gering zu halten und weitere Einflussvariablen auszuschließen. Die Voraussetzungen für ein Youden-Square-Design sind annähernd erfüllt (vgl. Frey, 2009): Die gebildeten Cluster kommen in den Testheften gleich häufig vor. Alle eingesetzten Aufgaben tragen zur Verankerung bei, d.h. eine Verlinkung zwischen den Testheften ist gewährleistet. Alle Testhefte weisen die gleiche Anzahl an Positionen auf, alle Kombinationen von jeweils zwei Clustern innerhalb eines Testhefts erscheinen in allen Testheften ebenfalls gleich häufig. Des Weiteren ist die Interaktion aus Clustern und Positionen kontrolliert. Auf die Bedingung, dass in jedem Testheft $n-1$ Cluster vorkommen, wurde hingegen verzichtet. Grund dafür waren die Voraussetzungen für das mögliche YSD aus den Parametern $t=b=12$ und $r=k=3$. Für diese Kombination existiert nur ein mögliches YSD mit 11 Clustern pro Testheft, aufgrund der hohen Bearbeitungszeit offener Aufgaben wurden insgesamt jedoch nur drei Positionen pro Testheft bestimmt. Trotzdem wurde sich für ein unvollständiges Blockdesign entschieden, das sich an einem Youden-Square-Design orientiert, da eine vollständige Permutation aus zeitlichen und Gründen der kognitiven Belastung nicht realisierbar gewesen wäre. Dennoch konnte so eine unerwünschte Variation der interessierenden Parameterschätzer kontrolliert werden (vgl. Frey, 2009). Für die Hauptstudie folgt daraus,

dass in 12 Testheften insgesamt auch 12 Cluster bestimmt werden konnten, wobei drei Cluster in jedem Testheft repräsentiert wurden und jeder Cluster in insgesamt drei Testheften und an jeder Position einmal vertreten ist (vgl. TABELLE 19, s. Kapitel 7.2). Die Verteilung der Testhefte und Cluster ist der TABELLE 22 zu entnehmen.

TABELLE 22: HAUPTSTUDIE – TESTHEFTDESIGN (36 ITEMS)

TH 1	TH 2	TH 3	TH 4	TH 5	TH 6
1	2	3	4	5	6
2	3	4	5	6	7
3	4	5	6	7	8
TH 7	TH 8	TH 9	TH 10	TH 11	TH 12
7	8	9	10	11	12
8	9	10	11	12	1
9	10	11	12	1	2

8 STATISTISCHE METHODEN DER HAUPTSTUDIE

Im Folgenden werden nur Methoden vorgestellt, die nicht bereits im Methodenteil zur Pilotierung (s. Kapitel 5) beschrieben wurden und für die Auswertung der Hauptstudie von besonderer Bedeutung sind.

8.1 PROBABILISTISCHE TESTTHEORIE FÜR MEHRDIMENSIONALE MODELLE

8.1.1 DAS MEHRDIMENSIONALE RASCH-MODELL

Für die Messung unterschiedlicher Komponenten einer Kompetenz werden mehrdimensionale Rasch-Modelle verwendet. Das mehrdimensionale Rasch-Modell basiert auf einem Itemkomponenten-Modell im Sinne eines Fähigkeitsmodells (vgl. Rost, 2004). Es handelt sich um hypothesentestende Modelle, da die Itemzuordnung zu den latenten Dimensionen bereits vor der Untersuchung festgelegt ist und die Modellanwendung prüft, ob die angenommene mehrdimensionale Teststruktur gültig ist (vgl. Rost, 2004). Die Modellgleichung für ein mehrdimensionales Rasch-Modell lautet:

$$P(x_{vi} = 1) = \frac{e^{\left(\sum_{j=1}^h q_{ij}\theta_{vj} - \sigma_i\right)}}{1 + e^{\left(\sum_{j=1}^h q_{ij}\theta_{vj} - \sigma_i\right)}}$$

Die Matrix mit den q_{ij} -Gewichten entspricht einer Q-Matrix für linear-logistische Testmodelle (Rost, 2004). Damit werden die Gewichte spezifiziert, mit denen eine latente Dimension j an der Lösung eines Items i beteiligt ist. Für die Modellierung von Kompetenzen und Kompetenzniveaus erscheinen mehrdimensionale IRT-Modelle (MIRT-Modelle) besonders vielversprechend (vgl. Frey & Hartig, 2009; Hartig & Jude, 2007). Über MIRT-Modelle lassen sich Items zu mehreren verschiedenen Fähigkeitskomponenten zuordnen. Die Wahrscheinlichkeit, ein Item richtig zu lösen, hängt in mehrdimensionalen Modellen von einem Vektor aus mehreren Fähigkeitsparametern ab und nicht wie in eindimensionalen Rasch-Modellen von genau einer latenten Fähigkeitsvariable (Reckase, 2009; Hartig & Höhler, 2010). Damit kann untersucht werden, wie viele verschiedene Fähigkeiten notwendig sind, um ein Item zu lösen. Dies erlaubt wiederum differenziertere Aussagen über vorhandene Kompetenzen und Teilkompetenzen der untersuchten Personen. Die Auswertung erfolgt dabei mehrdimensional, d.h. Teilkompetenzen werden als separate Dimensionen modelliert (Hartig & Jude, 2007). Dabei bezieht sich Mehrdimensionalität auf die Rasch-Modellierung und darf nicht mit dem Dimensionsbegriff, der häufig für Kompetenzmodelle verwendet wird, verwechselt werden. Eine Dimension im mehrdimensionalen Rasch-Modell entspricht je nach Modellierung einer

Komponente oder Unterkomponente im Kompetenzstrukturmodell. Aus testtheoretischer Sicht bedeutet dies, dass mit mehrdimensionalen Modellen mehrere latente Variablen, d.h. die in einem Test erfasste Kompetenz wird in Teilkompetenzen differenziert, modelliert werden können (Hartig & Klieme, 2006; vgl. Ackerman, Gierl, & Walker, 2003, Wang, 1997). Die Antwort auf die Frage nach einem mehrdimensionalen Konstrukt bleibt dennoch eine empirische. Je nach Passung unterschiedlicher statistischer Modelle auf die Daten, liegt Mehrdimensionalität vor oder nicht (Wilson, 2005). MIRT-Modelle bieten hingegen die Möglichkeit, Zusammenhänge zwischen verschiedenen Kompetenzen zu untersuchen, die jeweils durch eine separate latente Dimension modelliert werden. So kann in der vorliegenden Untersuchung die Repräsentationskompetenz, d.h. die Integration von MER durch die drei postulierten Teilkompetenzen *MERI*, *MERII*, *MERIII*, gemeinsam mit den zwei anderen Dimensionen, Aufgabenkontext und Repräsentationstyp, mehrdimensional modelliert und sensitiver analysiert werden (vgl. Adams, 2005; Adams & Wu, 2002).

Zur Messung von Kompetenzen wird für jede zu erfassende Kompetenz jeweils eine separate Dimension, d.h. latente Variable, definiert (Hartig & Höhler, 2010). Daneben können wiederum auch Subdimensionen innerhalb einzelner Kompetenzbereiche mehrdimensional modelliert werden. In beiden Fällen lädt jedes Item nur auf einer Dimension, d.h. die Lösung eines Items hängt genau von einer der betrachteten Fähigkeiten ab. Man spricht in diesem Zusammenhang von *Einfachstruktur* oder *Between-Item-Mehrdimensionalität* (vgl. Adams, Wilson, & Wang, 1997; Hartig & Höhler, 2010). Dabei werden sämtliche Items a priori zu disjunkten Gruppen zusammengefasst und werden den postulierten latenten Dimensionen zugeordnet. Die schematische Darstellung eines Between-Item-Modells kann ABBILDUNG 23A entnommen werden. MIRT-Modelle können auch eine komplexe Ladungsstruktur aufweisen. So werden Items, die als *Within-Item-Mehrdimensionalität* modelliert werden, gleichzeitig durch mehrere latente Dimensionen beeinflusst (vgl. ABBILDUNG 23B). Das heißt jedes Item kann einer oder mehreren Dimensionen zugeordnet werden. Damit erreicht man eine Modellierung, bei der zur Lösung eines Items eine oder mehrere Fähigkeitskomponenten benötigt werden (Hartig & Höhler, 2010; Wu et al., 2007). Die Anwendung dieser Between- und Within-Item-Mehrdimensionalität auf das postulierte Kompetenzmodell ist in Kapitel 8.1.3 dargestellt.

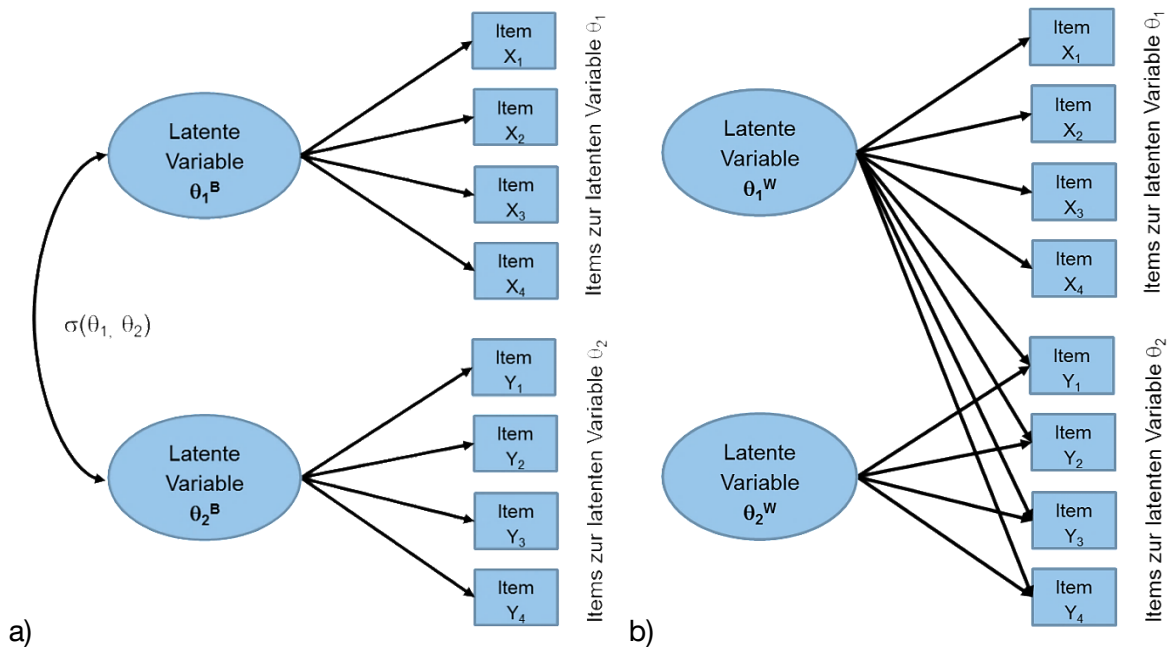


ABBILDUNG 23: SCHEMATISCHE DARSTELLUNG DER MODELLIERUNG VON TESTAUFGABEN X UND Y MIT A) ZWEIDIMENSIONALES IRT-MODELL MIT BETWEEN-ITEM-STRUKTUR UND B) ZWEIDIMENSIONALES IRT-MODELL MIT WITHIN-ITEM-STRUKTUR (ANGELEHNT AN HARTIG & HÖHLER, 2010)

Die Entscheidung für ein entsprechendes Modell sollte dabei nicht ausschließlich aus den empirischen Daten erfolgen. Die inhaltliche Bedeutung der latenten Variablen muss damit ausführlich dokumentiert und diskutiert werden (vgl. Hartig & Höhler, 2010). Generell erlauben MIRT-Modelle mit Mehrfachladungen (Within-Item-Mehrdimensionalität) Aussagen über mögliche Interaktionen zwischen den verschiedenen Teilkompetenzen und den Aufgabenanforderungen (Hartig & Höhler, 2010). Für die statistische Überprüfung, wie viele Dimensionen unterscheidbar sind, werden die Modellgütekriterien der verschiedenen postulierten Modelle miteinander verglichen. Des Weiteren werden die Korrelationen zwischen den Dimensionen analysiert und beurteilt (vgl. Hartig, 2007). Rost (2004) weist darauf hin, dass die Modellgeltungstests mit Likelihoodquotienten und mit den Informationskriterien AIC und BIC für mehrdimensionale Rasch-Modelle (ab zwei Dimensionen) eher problematisch anzusehen sind. Grund hierfür ist die starke Gewichtung der Parameteranzahl (v.a. beim BIC), die bei mehrdimensionalen Modellen ansteigt und aus der übergroßen Anzahl an Scoreparametern resultiert. Als Alternative schlägt Rost (2004) die Modellierung der Personenparameter, d.h. die Verteilung der latenten Variablen, vor. Das Verfahren gehört zur Maximum-Likelihood Methode, reduziert die große Anzahl multivariater Parameter und die Korrelationen der latenten Dimensionen sind zudem messfehlerbereinigt (Rost, 2004).

8.1.2 SCHÄTZUNG VON ITEM- UND PERSONENPARAMETERN

Im Folgenden wird an die bereits in Kapitel 5.2.2 beschriebenen Hintergründe zur Schätzung von Item- und Personenparameter angeknüpft und ausschließlich Erweiterungen, die für die Auswertung der Hauptstudie von Bedeutung sind, erläutert. Für alle ein- bis dreidimensionalen Rasch-Skalierungen wird das Gauß-Hermite-Quadratur-Verfahren angewendet. Für Modelle ab drei Dimensionen wird die Nutzung des Monte-Carlo-Verfahrens empfohlen und verwendet (Wu et al., 2007).

Bei mehrdimensionalen Konstrukten wird die Reliabilität für jede einzelne Dimension separat betrachtet. Die EAP/PV-Reliabilitäten können damit für Untersuchungen herangezogen werden, die Fähigkeitsverteilungen in Gesamtpopulationen beschreiben (Hartig & Kühnbach, 2006; Pohl & Carstensen, 2012). Im Rahmen der Hauptstudie werden Plausible Values (PVs) ebenfalls als Schätzer in die Testmodellierung aufgenommen, da für die Bestimmung der individuellen Fähigkeiten nur wenige Items pro Person eingesetzt wurden und sich PVs für eine unverzerrte Beschreibung von Populationen eignen (Wu, 2005). PVs sind logarithmisch transformierte Zufallswerte der a posteriori Verteilung der individuellen Messwerte (Carstensen et al., 2007; Rost, 2004, S. 328). Die Berechnungen erfolgen mit mehreren PVs, in der Regel werden fünf PVs gezogen, wobei meist ein PV pro Testperson ausreicht, um die Populationsparameter adäquat zu ermitteln (Wu, 2005). Berichtet wird dabei der erste PV (vgl. Baumert et al., 2000). Die PVs liegen als Messwerte auf der Logit-Skala, die als lineare Funktion von Personenfähigkeit und Itemschwierigkeit definiert wird (Rost, 2004). Typischerweise resultiert als Normbereich der Bereich zwischen -3 und +3 Logits (Rauch & Hartig, 2008). In diesem Zusammenhang stellen PVs die zuverlässigsten Schätzer dar, da Populationskennwerte bereits vor der Analyse in die Berechnung einfließen (Wu, 2005).

Die Schätzung der PVs erfolgt auf Basis der ermittelten Itemparameter. Dabei werden Personenvariablen berücksichtigt, die als Hintergrundmodell in Form von Regressionen in die Schätzung eingehen. Je mehr zur Verfügung stehende leistungsrelevante Informationen über eine Person, desto größer ist die Messgenauigkeit der Schätzung der latenten Fähigkeit (Carstensen et al., 2007; Rost, 2004). Fehlende Werte können dazu führen, dass der Regressionskoeffizient unterschätzt wird (Wu, 2005). Für eine präzisere Schätzung werden im Rahmen der Hauptstudie folgende Hintergrundvariablen in die Berechnung aufgenommen: 1. Studiengang, 2. Biologienote, 3. Mathematiknote und 4. das Geschlecht. Zusammenhänge zwischen Personenvariablen und Antwortverhalten können damit vor Schätzung der individuellen Fähigkeitsparameter zur Vorhersage der Fähigkeiten herangezogen werden.

Die Regressionen werden auf latenter Ebene modelliert und sind messfehlerbereinigt (Hartig & Kühnbach, 2007; Rost, 2004). Die so ermittelten unstandardisierten Regressionsgewichte dienen als Maß für den Einfluss einer Variablen auf das zu untersuchende Konstrukt.

8.1.3 MESSMODELLE DER UNTERSUCHUNG

Im Folgenden werden die Überlegungen zur differenzierten Beschreibung der Modellstruktur vorgestellt. Die Modellannahmen basieren auf der mehrdimensionalen Testanalyse nach Rost (2004, S. 259 ff.). Es wurde postuliert (s. Kapitel III), dass die Menge der konstruierten Testitems und deren Beantwortung von mehreren latenten Variablen abhängt, d.h. die Items sprechen mehrere Dimensionen an. Diese Annahmen sind aus der klassischen Testtheorie zur Faktorenanalyse bekannt (vgl. Rost, 2004). Mit Bezug zur Forschungsfrage eins (s. Kapitel III) werden folgende Experimental- und Kontrollmodelle auf ihre empirische Gültigkeit untersucht.

Das übergeordnete Experimentalmodell bezieht sich auf die Integration von MER (vgl. ABILDUNG 24). Auf Basis des beschriebenen Kompetenzmodells (s. Kapitel 3.3) werden die folgenden drei Dimensionen zum Umgang mit multiplen externen Repräsentationen angenommen: *Informationsentnahmenähe Integration (MERI)*, *Konstruktionsnahe Integration (MERII)* sowie *Translation und Transformation aus mehreren Repräsentationen (MERIII)*. In diesem Messmodell werden drei unterschiedliche Fähigkeiten als drei latente Variablen repräsentiert, d.h. empirisch erfolgt die beste Passung bei einem dreidimensionalen Modell. Dabei lassen sich die Testitems ohne Überschneidung auf jeweils eine dieser Fähigkeiten zuordnen (vgl. Hypothese $H_{1.1}$).

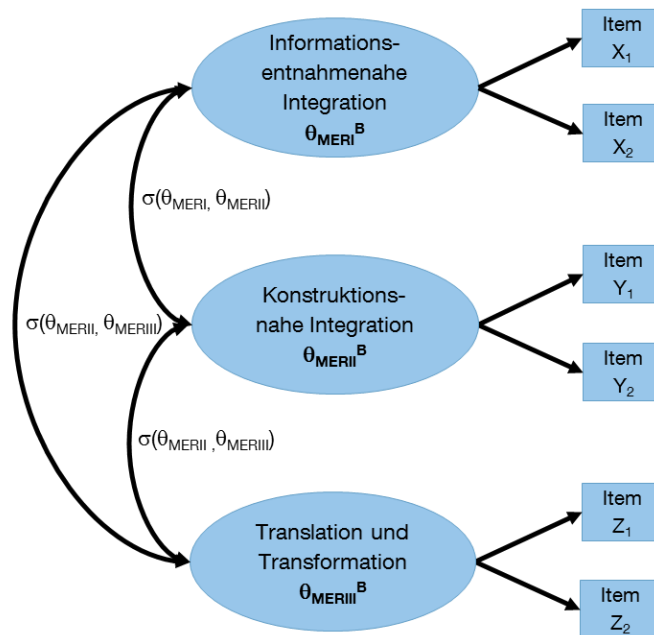


ABBILDUNG 24: 3-DIMENSIONALES MODELL (INTEGRATION VON MER ALS MERI, MERII, MERIII), BETWEEN-ITEM-MEHRDIMENSIONALITÄT UND ZU PRÜFENDER KORRELATIONEN ZWISCHEN DEN LATENTEN VARIABLEN

Im eindimensionalen Kontrollmodell (vgl. ABBILDUNG 25) wird angenommen, dass sich die Kompetenz *Integration von MER* durch genau eine Dimension abbilden lässt. Es lassen sich keine einzelnen Fähigkeiten differenzieren (vgl. Hypothese $H_{1.0}$).

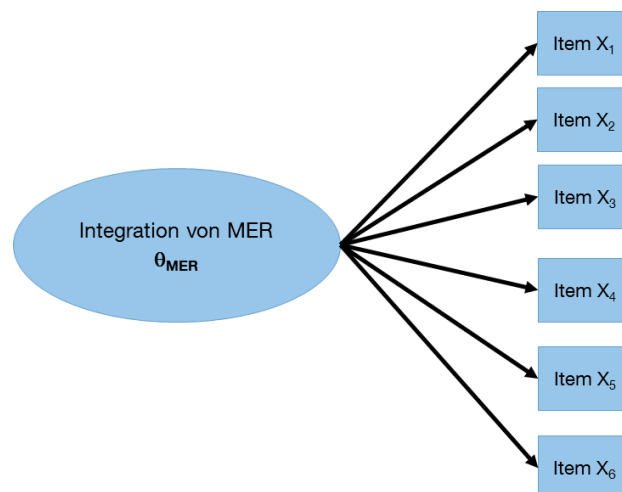


ABBILDUNG 25: 1-DIMENSIONALES MODELL ZUR INTEGRATION VON MER

Im Folgenden werden die Modelle vorgestellt, die im Rahmen der Integration von MER keine Teilkompetenzen unterscheiden, sondern die Integration vor dem Hintergrund des Aufgabenkontextes sowie des Repräsentationstyps darstellen (vgl. ABBILDUNG 26). Für beide Messmodelle liegt eine Between-Item-Mehrdimensionalität vor.

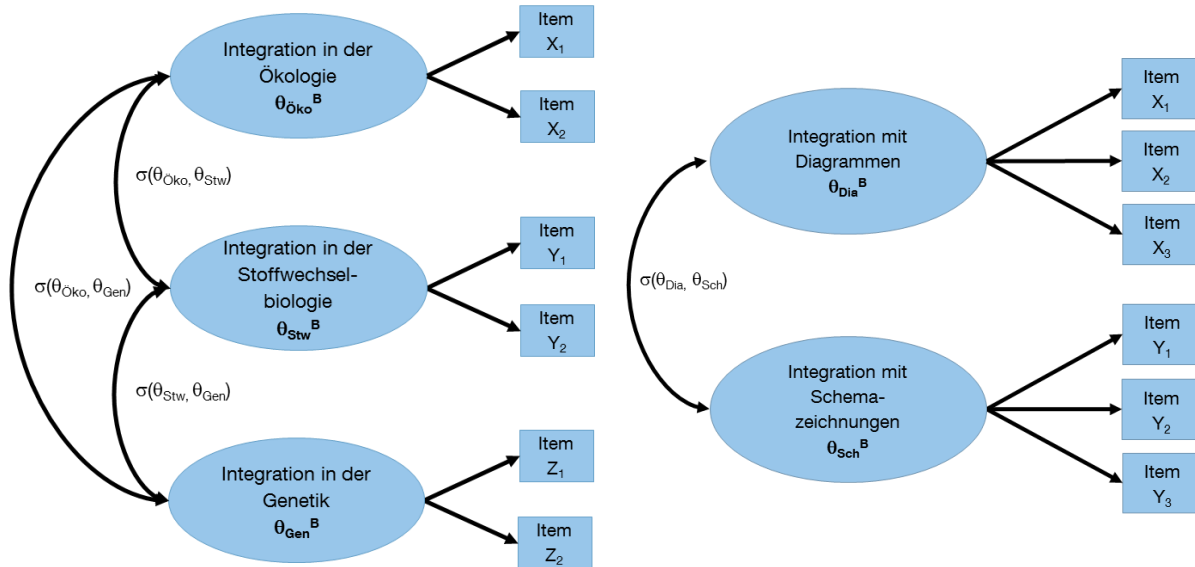


ABBILDUNG 26: 3-DIMENSIONALES (INTEGRATION NACH AUFGABENKONTEXT) UND 2-DIMENSIONALES MODELL (INTEGRATION NACH REPRÄSENTATIONSTYP), BETWEEN-ITEM-MEHRDIMENSIONALITÄT

Vor diesem Hintergrund sowie aufgrund der Konstruktion der Testaufgaben, die alle drei postulierten Dimensionen (Integration von MER, Aufgabenkontext, Repräsentationstyp) berücksichtigen, kann ein Experimentalmodell mit Within-Item-Mehrdimensionalität angenommen werden. In ABBILDUNG 27 sind verschiedene Kombinationsmöglichkeiten dargestellt, wodurch der Fall modelliert wird, dass zur Aufgabenlösung des jeweiligen Testitems eine oder mehrere Fähigkeitskomponenten benötigt werden.

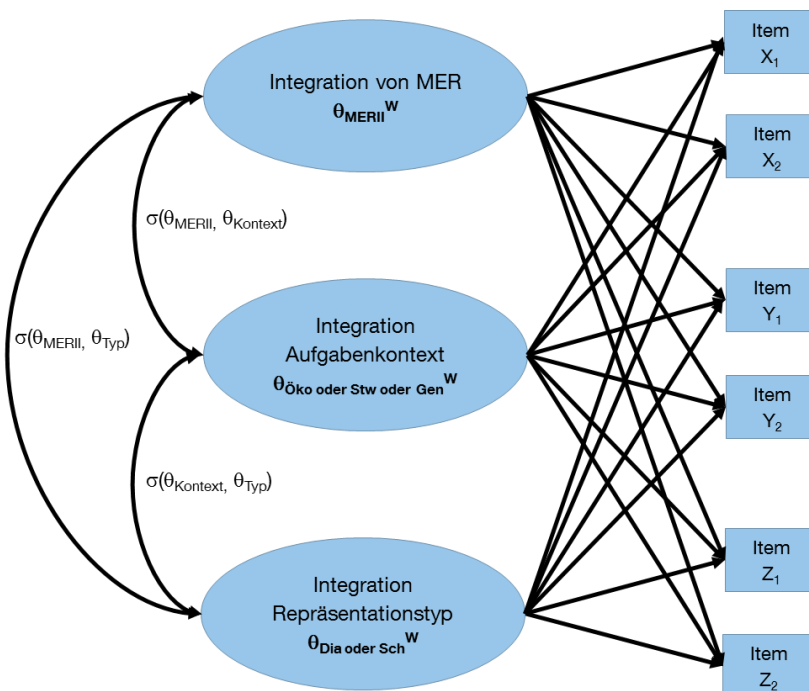


ABBILDUNG 27: MEHRDIMENSIONALES MODELL MIT UNTERSCHIEDLICHEN KOMBINATIONSMÖGLICHKEITEN ZUR WITHIN-ITEM-MEHRDIMENSIONALITÄT

In diesem Modell liegt eine Within-Item-Mehrdimensionalität vor. Jedes Testitem wird dabei jeweils einer der drei Dimensionen sowie ihrer Teilkomponente zugewiesen. Je nach Dimensionalität können hier weitere Dimensionsanalysen durchgeführt werden. So lässt sich annehmen, dass die Kompetenz *Integration von MER* durch die Teilkomponenten MERI, MERII, MERIII oder durch die Teilkomponenten MERI, MERII sowie die Teilkomponenten Ökologie, Stoffwechsel, Genetik beschrieben wird. Empirisch erfolgt demnach die beste Passung bei einem sechs- bzw. fünfdimensionalen Modell mit Within-Item-Mehrdimensionalität. Analog dazu gilt es zu prüfen, inwiefern ein fünf- bzw. vierdimensionales Modell, resultierend aus der Integration von MER als MERI, MERII, MERIII bzw. MERI, MERII sowie dem Repräsentationstypen Diagramm und Schemazeichnung, empirisch bestätigt werden kann. In diesem Zusammenhang sollte beachtet werden, dass für die Modellierung derartiger Modelle eine andere Interpretation der Messergebnisse zugrunde liegt, die zum Teil aus komplexen Modellierungen resultieren. So hängen bei Modellen mit Mehrfachladungen die Lösungswahrscheinlichkeiten der einzelnen Items von mehreren Fähigkeitsdimensionen ab (Hartig & Höhler, 2010). Aus diesem Grund muss geprüft werden, inwieweit diese postulierten Messmodelle mit den vorliegenden Items im Rahmen der Hauptstudie überhaupt durchgeführt und berechnet werden können.

Das postulierte Kompetenzmodell zum Umgang mit MER (vgl. ABBILDUNG 7) versucht neben den aus anderen Untersuchungen zur Diagrammkompetenz (vgl. Lachmayer, 2008; von Kotzebue, 2014) empirisch validierten Teilkompetenzen *Informationsentnahmenähe Integration (MERI)* und *Konstruktionsnähe Integration (MERII)* die dritte Dimension *Translation und Transformation aus mehreren Repräsentationen (MERIII)* auf ihre empirische Gültigkeit zu überprüfen. Die konstruierten Testaufgaben zur Teilkomponente MERIII enthalten dabei Anforderungen, die einerseits der Teilkomponente MERI, d.h. Translation als Informationsentnahmenähe Integration und andererseits eher der Teilkomponente MERII, d.h. Transformation als Konstruktionsnähe Integration, zugeordnet werden können. Aus diesem Grund wird eine zweidimensionale Betrachtung der MER-Integration ebenfalls untersucht. Die Zuordnung bezieht sich also auf den Output der Aufgabenlösung, d.h. einmal resultiert als Output ein Text (MERI) und zum anderen als Output ein Bild (MERII). Damit lassen sich die Testitems zu genau zwei disjunkten Skalen anordnen, indem die Items der MERIII-Integration eindeutig auf eine der beiden Teilkomponenten MERI oder MERII zugeteilt werden (Between-Item-Mehrdimensionalität). Vor diesem Hintergrund wurden die Ökologie- und Stoffwechsel-MERI-Items auf die Dimension MERI und die drei Genetik-MERIII-Items Blu08, Erb05 und Pku07 der Dimension MERII zugeordnet. Die Kontextabhängigkeit ist dabei rein zufällig. Für das zu prüfende Messmodell heißt das, dass die *Integration von MER* um

die Dimension MERIII reduziert wird und nur durch die zwei Teilkomponenten *Informationsentnahmensnahe Integration (MERI)* und *Konstruktionsnahe Integration (MERII)* abgebildet wird (vgl. ABBILDUNG 28).

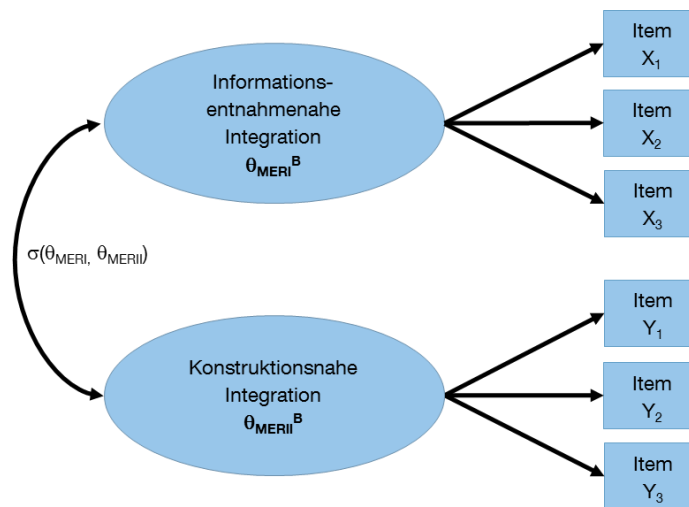


ABBILDUNG 28: 2-DIMENSIONALES MODELL ZUR INTEGRATION VON MER (MERI, MERII), BETWEEN-ITEM-MEHRDIMENSIONALITÄT

8.2 PRÜFUNG DER MODELLPASSUNG

Für die postulierte Kompetenzstruktur werden verschiedene Rasch-Modelle modelliert (s. Kapitel 8.1.3). Die Überprüfung der Modellpassung erfolgt dabei in zwei Schritten: 1. Der Reliabilitätsprüfung und 2. die Prüfung der Modellpassung und Modelldimensionen.

8.2.1 PRÜFUNG DER RELIABILITÄT

In einem ersten Schritt wird die Zuverlässigkeit der Schätzung für die Personen und Itemparameter in jedem Modell anhand von Kennwerten bewertet. Das Programm ConQuest gibt dabei die *Person Separation Reliability* und die *Item Separation Reliability* aus. Die empirischen Werte werden dabei zu den in der Literatur beschriebenen Grenzwerten in Beziehung gesetzt. Beide Maße liegen in einem Wertebereich zwischen 0 und 1 (Bond & Fox, 2007). Die *Person Separation Reliability* ist ein Maß für die Differenzierbarkeit der Personen bezüglich ihrer Merkmalsausprägung. Je höher diese ist, desto eher können Probanden mit Personenparameter am oberen Ende der Skala auch tatsächlich eine höhere Merkmalsausprägung besitzen als solche Personen, deren Personenparameter am Ende der Skala zu finden sind. In ConQuest können verschiedene Varianten der *Person Separation Reliability* angefordert werden (Wu et al., 2007). Neben der MLE Person Separation Reliabilität können die WLE Person Separation Reliabilität und die EAP/PV Person Separation Reliabilität herangezogen

werden. Rost (2004) berichtet, dass die WLE Person Separation Reliability bei vielen Personen mit Extremscores in der Stichprobe eher unterschätzt wird, weshalb bevorzugt die EAP/PV Person Separation Reliability angegeben wird. Gleichzeitig zeigt dieses Reliabilitätsmaß eine große Übereinstimmung mit *Cronbachs Alpha* aus der klassischen Testtheorie auf und wird aus diesem Grund auch in der vorliegenden Arbeit angegeben (Rost, 2004). Die Item Separation Reliability ist ein Maß für die Differenzierbarkeit der ermittelten Itemschwierigkeiten. Sie ist typischerweise sehr hoch und steigt bei zunehmender Stichprobengröße an (Wu et al., 2007).

8.2.2 PRÜFUNG DER MODELLGÜLTIGKEIT UND MODELLEDIMENSIONEN

Bevor eine Interpretation des Kompetenztests vorgenommen wird, muss die Modellpassung überprüft werden, d.h. wie gut das angewendete Modell zu den empirischen Daten passt (Rost, 2004). In einem ersten Schritt wird geprüft, ob die Anpassung des Modells erfolgreich und innerhalb der maximalen Zahl von Iterationen abgeschlossen wurde (Neumann, 2014). Anschließend folgt die Überprüfung der Modellgültigkeit anhand verschiedener Kennwerte.

Bei einer eindimensionalen Skalierung und der Einstellung „*constraints = cases*“, werden nicht die Item- sondern die Personenparameter summennormiert. Damit werden alle Itemparameter (Itemschwierigkeitsmittel und Standardabweichung) frei geschätzt (Wu et al., 2007). Wird ein eindimensionales Rasch-Modell unterstellt, wird für die Lösung der Aufgaben die Beherrschung genau eines Fähigkeitskonstrukts angenommen (Kauertz, 2014). Ziel einer wissenschaftlichen Modellierung ist die Beschreibung eines Sachverhalts mit möglichst wenigen Annahmen (Parametern), was unter dem Begriff des „Einfachheitskriteriums“ zusammengefasst wird (vgl. Rost, 2004). Werden verschiedene Modelle postuliert, muss zur Beurteilung der Modellgüte abgewogen werden, wie gut das Modell auf die empirischen Daten passt und mit vielen Parametern diese Passung erzielt wurde. Dabei gilt es zu beachten, dass keine globalen Modellgütemaße geprüft werden können, d.h. die Richtigkeit eines Modells kann aus wissenschaftstheoretischer Sicht nicht endgültig bewiesen werden (Bühner, 2006). Aus diesem Grund müssen die postulierten Modellannahmen überprüft und eine Aussage, welches dieser Modelle die Daten am besten beschreibt, getroffen werden (Rost, 2004). Zur Interpretation werden verschiedene Informationskriterien herangezogen sowie die „Sparsamkeit“ der verschiedenen Modelle miteinander verglichen (Bühner, 2006). Im Rahmen der vorliegenden Untersuchung werden drei informationstheoretische Indices herangezogen, die sowohl die log-Likelihood (logarithmierte Likelihoodwerte) als Anpassungsmaß und die Zahl der Modellparameter n_p berücksichtigen (Rost, 2004). In beide Verfahren geht in die Berechnung des Model-Fit die Final Deviance (Maß für die absolute Abweichung des Modells von

den empirischen Daten, negativ doppelter log-Likelihood) und die Parameterzahl ein. Die Anzahl der Parameter gibt an, wie viele Annahmen getroffen werden müssen, damit die Passung erreicht wird (vgl. Ziepprecht, 2016, S. 127). Je kleiner die Final Deviance und die Parameteranzahl, desto sparsamer ist das Modell und desto besser passt das Modell. Über die Differenz der Final Deviance-Maße im Verhältnis zu den Freiheitsgraden (Differenz der Parameterzahl der beiden zu vergleichenden Modelle) kann der Unterschied im Model-Fit unter Berücksichtigung einer akzeptablen Irrtumswahrscheinlichkeit geprüft werden (Senkbeil et al., 2005). Hierfür existieren drei informationstheoretische Maße: Akaike Information Criterion (AIC), Best Information Criterion (BIC) und das Consistent Akaike Information Criterion (CAIC). Für alle Indices gilt: Je geringer der Index, desto besser passt das postulierte Modell zu den empirischen Daten (Rost, 2004). Neben der Final Deviance gehen in alle drei Indices auch die Modellkomplexität (Modellparameter n_p) und die Stichprobengröße (N) in unterschiedlicher Gewichtung ein (Bühner, 2011, S. 542). Das AIC gewichtet die Parameterzahl n_p und die Final Deviance (als negativ doppelter log-Likelihood) im Verhältnis 1 zu 1 (Rost, 2004):

$$AIC = -2 \log L + 2 n_p$$

Das AIC eignet sich insbesondere für große Stichproben. Burnham & Anderson (2004) berichten, dass der AIC vor allem dann herangezogen werden soll, wenn das Verhältnis zwischen Stichprobengröße und Parameterzahl aller miteinander verglichenen Modelle geringer als 40 ist. Das BIC gewichtet die Parameterzahl höher, d.h. dass ein komplexeres Modell nach dem AIC eine bessere Passung, aber nach dem BIC eine schlechtere Passung aufweisen kann als ein sparsames Modell. Dem Einfachheitskriterium wird dadurch eine höhere Bedeutung zugemessen (vgl. Ziepprecht, 2016, S. 127). Das BIC gewichtet die Parameterzahl mit einem variablen Koeffizienten, dem Logarithmus der Stichprobengröße N (Rost, 2004):

$$BIC = -2 \log L + (\log N) n_p$$

Das BIC eignet sich damit für große Itemzahlen und für die Anwendung unterschiedlicher Antwortmuster (vgl. Roeling, 2016, S. 77), während das Kriterium weniger bei großen Stichproben geeignet ist (Kuha, 2004; Posada & Buckley, 2004). Aus dem AIC leitet sich der dritte Index ab, das CAIC. Dabei handelt es sich um einen korrigierten AIC, der auch bei großen Stichproben konsistent ist (Rost, 2004).

$$CAIC = -2 \log L + (\log N) n_p + n_p$$

Hinsichtlich der Interpretation der drei unterschiedlichen Indices können folgende Annahmen getroffen werden. Das Modell für das kleinste informationstheoretische Maß passt im Vergleich zu allen anderen Modellen am besten zu den Daten (Rost, 2004). Falls AIC und BIC für dasselbe Modell sprechen, gilt dieses Modell, da es „unlikely to be far from the best of the candidate models [ist]“ (Kuha, 2004, S. 223). Ist die Differenz zwischen zwei Indices größer als zehn, liegt eine bessere Passung des Modells zu den Daten mit dem niedrigeren informationstheoretischen Maß vor (Burnham & Anderson, 2004).

Neben dem bloßen Vergleich in den Unterschieden der Final Deviance kann zusätzlich ein *Likelihoodquotiententest* durchgeführt werden. In diesem Verfahren erfolgt ein Vergleich der Likelihoods von verschiedenen Modellen mit derselben Datenmatrix. Dabei müssen drei Bedingungen erfüllt sein (vgl. Rost, 2004). Das Modell, dessen Likelihood im Nenner steht, muss ein echtes Obermodell sein, das restriktive Alternativmodell darf nicht durch Null-Setzen einzelner Parameter aus dem allgemeineren Modell resultieren und für das allgemeinere Modell muss im Nenner die Modellgültigkeit nachgewiesen sein. Sind diese Bedingungen erfüllt, wird der Likelihoodquotient in eine χ^2 -Prüfstatistik umgewandelt (ebd.). Der Likelihoodquotient $-2 \log \left(\frac{L_0}{L_1} \right)$ berechnet sich dabei aus der Differenz der Devianzmaße (ConQuest: Final Deviance) von zwei Modellen, die zu der Anzahl der Freiheitsgrade (Differenz der Parameterzahl der beiden Modelle, ConQuest: total number of estimated parameters) in Beziehung gesetzt wird und anhand der χ^2 -Verteilung geprüft wird (Senkbeil et al., 2005). Wie in anderen Signifikanztests sollte auch die χ^2 -Verteilung ein Signifikanzniveau kleiner als $p = 0,05$ bzw. fünf Prozent aufweisen (Rost, 2004).

In ConQuest können zusätzlich die *Interkorrelationen* der postulierten Dimensionen auf latenter Ebene (vgl. ABBILDUNG 23, $\sigma(\theta_1, \theta_2)$) als Entscheidungskriterium herangezogen werden. Die Interkorrelationen werden als Modellparameter geschätzt und sind messfehlerbereinigt (Carstensen et al., 2007). Korrelieren die postulierten Konstrukte nur gering untereinander, ist es naheliegend von einem mehrdimensionalen Konstrukt auszugehen (vgl. Kauertz, 2014). Hohe Zusammenhänge mit Werten größer als 0,90 sprechen hingegen für die Eindimensionalität, Werte darunter für die Mehrdimensionalität des Modells (Bond & Fox, 2007; Hartig & Klieme, 2006). Die Modellpassung wird nicht ausschließlich durch Kriterien der Psychometrie vorgenommen. Zusätzlich erfolgt eine Analyse der Korrelationen zwischen verschiedenen Aufgabenmerkmalen, um gegebenenfalls neue fachdidaktische Kenntnisse aufzudecken. Im Rahmen der vorliegenden Untersuchung können damit auch unterschiedliche Aufgabenmerkmale aus fachlicher Sicht, indem unterschiedliche Inhaltsbereiche der Biologie abgeprüft werden, gerechtfertigt werden. Die Feinstruktur ist dann nicht psychometrisch, sondern

aus fachdidaktischer Sicht sinnvoll (vgl. Kauertz, 2014). Liegt eine geringe Korrelation zwischen den Aufgabenmerkmalen vor, sollte von einem mehrdimensionalen Konstrukt ausgegangen werden, wozu weitere Rasch-Analysen benötigt werden.

8.3 WEITERE AUSWERTUNGSMETHODEN DER HAUPTSTUDIE

8.3.1 ANALYSE DER ITEM- UND PERSONENPARAMETER

Die Analyse der Itemparameter folgt den in Kapitel 5.3.3. genannten Kriterien. Ergänzend dazu wird in der Hauptstudie durch die Analyse der Aufgabenparameter geprüft, inwieweit sich der Test auf die getestete Stichprobe verhält (Neumann, 2014).

Im Statistikprogramm ConQuest kann mittels der *Regression Coefficients* die Abhängigkeit der Personenfähigkeit von zuvor definierten Regressionsvariablen mit einer linearen Regression untersucht werden (Wu et al., 2007). Als Wert wird der *Constant*, die mittlere Personenfähigkeit (unter „*constraints = items*“), herangezogen. Liegt eine negative Personenfähigkeit vor, weist dies auf eine mittlere Personenfähigkeit hin, die kleiner als die mittlere Aufgabenschwierigkeit ist. Damit ist der Test für die untersuchte Stichprobe (etwas) zu schwer. Zusätzlich kann über die *Variance* die Varianz der Personenverteilung betrachtet werden. Erstrebenswert sind Werte größer als 1,0, da eine zu kleine Varianz auf eine mangelnde Unterscheidung zwischen den Testpersonen hindeutet (Neumann, 2014). Kauertz (2008) berichtet, dass hohe Werte in der Varianz vor allem erreicht werden können, wenn verschiedene Inhaltsbereiche getestet werden. Neumann (2014) weist darauf hin, dass bei zu kleinen Varianzen überprüft werden sollte, inwieweit der Test das Fähigkeitsspektrum der Testpersonen abdeckt. Zur Untersuchung des Fähigkeitsspektrums kann die *Wright Map*, die als Output direkt der statistischen Software entnommen ist, herangezogen werden. In dieser Map sind Personenfähigkeiten und Itemschwierigkeiten auf einer gemeinsamen Logit-Skala aufgetragen. Dabei ist auf eine gleichmäßige Abdeckung der Personenfähigkeiten zu den gelösten Aufgaben zu achten (Neumann, 2014). Große Lücken können die Reliabilität des Messinstruments beeinträchtigen, da sich eine größere Personengruppe nicht mehr unterscheiden lässt. Für die Analyse der latenten Fähigkeit kann die Wright Map inhaltlich interpretiert werden. Das Item, das in der Wright Map ganz unten zu finden ist, deutet auf ein basales Verständnis hin. Anschließend folgt man der Wright Map immer weiter nach oben und identifiziert Itemgruppen, die auf ein nächst höheres Verständnis hinweisen. Als Reliabilitätsmaß dient zum einen die *WLE Person Separation Reliabilität*. Zusätzlich wird die IRT-basierte EAP/PV-Reliabilität als Skalenhomogenitätsmaß herangezogen (Wu et al., 2007, s. Kapitel 8.2.1).

Zur Darstellung von Leistungsunterschieden in der Grundgesamtheit *Integration von MER* (MERI, MERII, MERIII) oder in der Grundgesamtheit *Aufgabenkontext* (Ökologie, Stoffwechsel, Genetik) bzw. *Repräsentationstyp* (Diagramm, Schemazeichnung), werden die Personenfähigkeiten analysiert. Dafür wird die abhängige Variable (Personenfähigkeit) in ConQuest für jede Dimension geschätzt und je nach Anzahl der Aspekte entweder drei (Integration von MER, Aufgabenkontext) oder zwei (Repräsentationstyp) Plausible Values (PVs) ermittelt. Die Mittelwerte der abhängigen Gruppen können mittels einfaktorieller ANOVA mit Messwiederholung verglichen werden (vgl. Wellnitz, 2012, S. 88). Voraussetzung ist, dass die Varianz der Differenzen (Sphärizität) zwischen den verschiedenen Variablen gleich ist beziehungsweise diese voneinander unabhängig sind. Bei großer Verletzung dieser Voraussetzung werden multivariate Verfahren wie z.B. MANOVA vorgezogen, bei kleiner Verletzung univariate Varianzanalysen unter Greenhouse-Geisser Korrektur (Field, 2009).

8.3.2 KORRELATIONSANALYSEN

Mittels manifester und latenter Korrelation werden Zusammenhangsanalysen durchgeführt. Liegen beide Variablen intervallskaliert und normalverteilt vor, können Produkt-Moment-Korrelationen nach Pearson (r) berechnet werden (Field, 2009). Als nicht-parametrisches Verfahren wird für Zusammenhangsanalysen bei mindestens einer ordinalskalierten Variable der Rang-Korrelationskoeffizient nach Spearman (ρ) herangezogen (Field, 2009). Beide Koeffizienten liegen zwischen den Werten -1 (negativer linearer Zusammenhang) und $+1$ (positiver linearer Zusammenhang), wobei Werte bis $0,10$ als kleine, bis $0,30$ als mittlere und Werte $\geq 0,50$ einen starken Zusammenhang signalisieren (Field, 2009; Bühner & Ziegler, 2009; Cohen, 1988). Bei $r = 0$ besteht kein Zusammenhang (Brosius, 2013; Bortz & Schuster, 2010). Für die vorliegende Untersuchung wird die Einstufung nach Zöfel (2003) herangezogen, da hier die Werte weiter differenziert werden (vgl. TABELLE 23).

TABELLE 23: EINSTUFUNG DES KORRELATIONSKOEFFIZIENTEN R (NACH ZÖFEL, 2003)

Korrelationskoeffizient	Einstufung
$ r \geq 0,2$	sehr geringe Korrelation
$0,2 < r \leq 0,5$	geringe Korrelation
$0,5 < r \leq 0,7$	mittlere Korrelation
$0,7 < r \leq 0,9$	hohe Korrelation
$0,9 < r \leq 1,0$	sehr hohe Korrelation

Im Rahmen dieser Arbeit werden Zusammenhänge z.B. zwischen den Variablen Personenfähigkeit, d.h. die Kompetenzen bei der *Integration von MER*, und der Schulnote in Biologie

und Mathematik berechnet. Dabei wurden die Noten in Punktwerten von Null bis 15 erfragt. Bei einem negativen linearen Zusammenhang bedeutet dies, dass je besser die Biologie- oder Mathematiknote, desto geringer sind die Kompetenzen bzw. bei einem positiv linearen Zusammenhang, je besser die Biologie- oder Mathematiknote, desto höher sind die Kompetenzen (Bortz, 2005). Die bivariate Korrelationsanalyse wird in IBM SPSS 22 durchgeführt. Korrelationsvergleiche werden angestellt, um signifikante Unterschiede zwischen verschiedenen Korrelationskoeffizienten zu ermitteln (Field, 2009). Das Quadrat R^2 des Korrelationskoeffizienten ist das Bestimmtheitsmaß B . Damit kann das Verhältnis von erklärter Varianz zur Gesamtvarianz interpretiert werden. Bei den Korrelationskoeffizienten unabhängiger Stichproben erfolgt eine Fisher-Transformation in Z-Werte, um mehrere Korrelationskoeffizienten zu mitteln (Bortz & Schuster, 2010). Dadurch können zusätzlich Verhältnisse interpretiert werden (Guilford & Fruchter, 1978, zitiert nach Bortz & Schuster, 2010, S. 161).

8.3.3 REGRESSIONSANALYSEN

Im Anschluss an die Korrelationsanalyse von zwei Merkmalen bzw. Variablen, kann das eine Merkmal zur Vorhersage des anderen eingesetzt werden (Bortz & Schuster, 2010). Das Ziel einer Regressionsanalyse besteht darin, die Unterschiede in einer Variable auf eine andere Variable zurückzuführen, während eine Korrelation lediglich den Zusammenhang zwischen zwei Variablen beschreibt (Bühner & Ziegler, 2009). Für diesen Unterschied wird eine Regressionsgleichung zwischen den zwei stochastisch abhängigen Variablen aufgestellt. Man unterscheidet dabei die *Prädiktorvariable*, d.h. die zur Vorhersage eingesetzte Variable (unabhängige Variable), und die *Kriteriumsvariable*, d.h. die Variable, die vorhergesagt werden soll (abhängige Variable) (Bortz & Schuster, 2010).

Damit lassen sich die empirisch ermittelten Aufgabenschwierigkeiten mit den zu analysierenden Aufgabenmerkmalen in Beziehung setzen (vgl. Hartig, 2007). Die Regressionsanalyse untersucht, ob die Unterschiede der Aufgabenschwierigkeiten durch die Verwendung der Aufgabenmerkmale erklärt werden können und welche Bedeutung bzw. Erklärungskraft diese Merkmale für die Schwierigkeit haben (Knigge, 2010). Eine Regression kann als lineares Modell aus den erhaltenen Messdaten ermittelt werden (Field, 2009). Die Schätzung für ein latentes Regressionsmodell erfolgt direkt innerhalb des IRT-Modells ohne Bezug zu den Personenschätzern (Adams & Wu, 2010). Die resultierenden Regressionsgewichte geben an, um wieviel sich die abhängige Variable ändert, wenn sich die unabhängige Variable um eine Einheit ändert (Field, 2009). Mittels eines linearen Regressionsmodells zwischen Aufgabenmerkmalen und Aufgabenschwierigkeiten erfolgt die Analyse für die verschiedenen Tests. Für lineare Regressionsmodelle gelten die Voraussetzungen Intervallskalenniveau der abhängigen

Variablen, eine Normalverteilung der Residuen², Linearität, Residuen-Varianzhomogenität und stochastische Unabhängigkeit der Residuen (Field, 2009). Nur dann kann von einem Signifikanztest ohne Verzerrung ausgegangen werden. Werden Merkmale mit mehr als zwei Ausprägungen untersucht, die kein Intervallskalenniveau aufweisen, müssen die einzelnen Merkmale in sogenannte Dummy-Variablen umgewandelt werden, die Werte von 1 und 0 annehmen (z.B. Brosius, 2013; Bühner & Ziegler, 2009; Bortz, 2005; Backhaus et al., 2003). Bei nominaler oder ordinaler Codierung der Variablen wie z.B. dem Geschlecht, wird eine solche Dummy-Codierung angewendet. Die Analysen können als multiple Regressionsmodelle nach Transformation der Daten aus dem Statistikprogramm ConQuest in SPSS umgesetzt werden. Stochastische Unabhängigkeit der Residuen bedeutet, dass diese zufällig auftreten und keinem bestimmten Muster folgen, da sonst nicht alle relevanten Prädiktorvariablen eingeschlossen werden (Brosius, 2013). Zusätzlich sollten die Residuen normalverteilt sein, was anhand von grafischen Verfahren (Histogramme, Normalverteilungsdiagramm) in SPSS überprüft wird (Field, 2009). Hinsichtlich der erklärenden Variablen (Prädiktorvariablen) gilt außerdem, dass diese nicht miteinander korrelieren dürfen, d.h. es darf keine Kollinearität vorliegen (Brosius, 2013). Die Überprüfung erfolgt mittels Korrelationsmatrix, ein Maß für Kollinearität ist der Variance Inflation Faktor (VIF), wobei Werte < 10 auf keine Kollinearität hinweisen (Field, 2009). Alternativ könnte eine Faktorenanalyse durchgeführt werden, um Multikollinearität zwischen den zu analysierenden Personenvariablen zu verhindern (Field, 2009). In diesem Fall können Faktorwerte generiert werden, die nicht miteinander korrelieren und zur Varianzaufklärung beitragen. Bei einer Regression darf zusätzlich keine Autokorrelation der Residuen vorliegen. Um diese auszuschließen, wird der Durbin-Watson-Koeffizient herangezogen, der Werte zwischen 0 und 4 annimmt. Je näher der Wert an 2 liegt, desto geringe kann von einer Autokorrelation der Residuen ausgegangen werden (Field, 2009).

Für die vorliegende Arbeit werden die genannten Voraussetzungen geprüft. Da die eingesetzten Personenparameter aus der Rasch-Analyse per Definition Intervallskalenniveau besitzen, gilt die Voraussetzung als erfüllt. Für die unabhängigen Variablen erfolgte eine Umcodierung der Prädiktorvariable in eine Dummy-Variable (Brosius, 2013). Dies trifft auf die Variable *Studiengang* zu, bei der unterschiedliche Kategorien wie beispielsweise Lehramt Gymnasium mit Fächerkombination Biologie unterschieden werden. Auch die Aufgabenmerkmale *Integration von MER*, *Aufgabenkontext* und *Repräsentationstyp* fließen als Prädiktoren für die Lösungswahrscheinlichkeit der Items und die Itemschwierigkeiten als Kriteriumsvariablen

² Als Residuen werden die Vorhersagefehler, d.h. die Differenzen zwischen den beobachteten Werten und dem aufgrund der Regressionsgeraden vorhergesagten Werten in der Stichprobe bezeichnet (Bortz & Schuster, 2010, S. 186).

(Logits) in die Berechnungen ein. Damit kann der Einfluss dieser Merkmale (unabhängige Variablen) auf die Itemschwierigkeit oder der Einfluss des besuchten Studiengangs auf die Kompetenzen einer Person mittels einer multiplen linearen Regression bestimmt werden. Für die Umcodierung heißt das, dass für jede Abstufung innerhalb einer Dimension bzw. zu jeder Kategorie eine neue Variable (Dummy-Variable) generiert wurde (vgl. Field, 2009; Prenzel et al., 2002). Anschließend wurde jeder Testperson, die die Kategorie bedient (z.B. den Studiengang Lehramt Gymnasium mit Fächerkombination Biologie) der Wert 1 zugewiesen. Alle anderen Testpersonen erhielten die Codierung 0. Bei allen weiteren Variablen wurde analog verfahren. Die Dummy-Codierungen werden als separate Merkmale behandelt und dafür jeweils eigene standardisierte Regressionsgewichte β geschätzt. Anschließend kann durch den Vergleich dieser standardisierten Regressionsgewichte der relative Einfluss der verschiedenen Prädiktoren auf die Aufgabenschwierigkeit ermittelt werden, indem aufgezeigt wird, ob ein Item mit jeweiliger Merkmalsausprägung schwerer als ein Item ist, bei dem dieses Merkmal nicht vorliegt (Hartig, 2007, zitiert nach Wellnitz, 2012, S. 87). Bei einem Regressionsgewicht von $\beta = 0$ besteht kein Zusammenhang zwischen Aufgabenmerkmal und Itemschwierigkeit, hohe β -Werte signalisieren hohe Zusammenhänge und ermöglichen damit eine größere Varianzaufklärung innerhalb der Schwierigkeit. Damit werden Aussagen möglich, welche unabhängige Variable wie stark zur Prädiktion beiträgt, da ein Vergleich der Regressionsgewichte mit unterschiedlicher Metrik möglich ist (Bühner, 2006). Mittels der unstandardisierten Regressionsgewichte (B) können die Aufgabenschwierigkeiten mit bestimmten Merkmalen vorhergesagt werden (Wellnitz, 2012). Dabei wird vorhergesagt, um wie viele Einheiten sich die abhängige Variable verändert, wenn sich die unabhängige Variable um eine Einheit verändert (Ziepprecht, 2016). Zusätzlich wird im Rahmen der Regressionsanalyse das Bestimmtheitsmaß R^2 ausgegeben. Der Determinationskoeffizient gibt Aufschluss wie sehr die Unterschiede in einem Prädiktor durch Unterschiede im anderen Prädiktor erklärt werden kann und nimmt Werte zwischen 0 und 1 an (Bühner & Ziegler, 2009). Bei der Berichterstattung wird über die F-Statistik die globale Passung der Regressionsgleichung angegeben, indem die Abweichungen der Modellgleichung von der „Nullfunktion“ ausgegeben werden. Der F-Test ist damit ein Signifikanztest, ob die angenommenen unabhängigen Variablen einen Beitrag zur Erklärung der Varianz liefern (Janssen & Laatz, 2005). Mittels korrigiertem Bestimmtheitsmaß R^2_{kor} wird aufgezeigt, wie gut die Modellfunktion die Daten erklärt. In diesem Zusammenhang wird zusätzlich die Effektgröße ε^2 herangezogen, um schwache, mittlere und starke Effekte berichten zu können (Döring & Bortz, 2016, S. 817 ff.; mod. nach Cohen, 1988). Für die Interpretation des Bestimmtheitsmaßes R^2_{kor} wird demnach ein Wert von $R^2_{\text{kor}} \leq 0,02$ als kleiner, Werte $R^2_{\text{kor}} \leq 0,13$ als mittlerer und Werte $R^2_{\text{kor}} \leq 0,26$ als großer Effekt klassifiziert (vgl. Döring & Bortz, 2016, S. 821). Bei einem statistisch signifikanten F-Wert wird

überprüft, inwiefern sich der Determinationskoeffizient R^2 eines jeden zu prüfenden Prädiktors signifikant von Null unterscheidet. Ein hohes Bestimmtheitsmaß signalisiert eine hohe Korrelation zwischen abhängiger und unabhängiger Variable, wodurch zuverlässige Rückschlüsse auf die abhängige Variable abgeleitet werden können (Brosius, 2013). Green (1991) nennt in diesem Zusammenhang bei Durchführung von multiplen Regressionen ein Minimum von $50 + 8k$ für den Determinationskoeffizienten R^2 und für die Regressionsgewichte β ein Minimum von $104 + k$ (k steht für die Anzahl der Prädiktoren).

9 ERGEBNISSE DER HAUPTSTUDIE

Anhand der erhobenen Hauptstudien­daten werden die in Kapitel III formulierten Hypothesen geprüft. Dabei wird zunächst die psychometrische Qualität der Testaufgaben analysiert und dokumentiert. Darauf basierend wird das postulierte Kompetenzstrukturmodell auf seine empirische Gültigkeit geprüft. Die angenommenen Dimensionen und ihre Teilkomponenten werden im Rahmen der Modellprüfung analysiert. In diesem Zusammenhang wird die interne Konsistenz der Skalen sowie die Trennschärfe der Testitems sichergestellt. Abschließend werden Unterschiedsanalysen im Hinblick auf die Itemschwierigkeit und Personenfähigkeit durchgeführt sowie untersucht, ob Zusammenhänge zwischen dem Integrationsprozess von multiplen externen Repräsentationen (MER) und anderen Variablen wie z.B. Studiengang, Schulnoten und Abitur bestehen. Eine qualitative Fehleranalyse der 36 beantworteten Testitems beendet das Kapitel. Die Fehleranalyse wird im Überblick an ausgewählten Items dargestellt und dabei auffällige Kriterien in der Bearbeitung identifiziert.

9.1 TESTITEMS UND MODELLPASSUNG

Im ersten Schritt werden die in der Hauptstudie eingesetzten 36 Testaufgaben (s. Kapitel 6.3) auf ihre psychometrische Qualität untersucht. Dafür wird für jedes der angenommenen ein- oder mehrdimensionalen Modelle (s. Kapitel 8.1.3) eine fallzentrierte Rasch-Skalierung durchgeführt. Als zentrale Indikatoren der Passung des Rasch-Modells zu den empirischen Daten werden die gewichteten Abweichungsquadrate zwischen erwarteten und beobachteten Häufigkeiten ($wMNSQ$) für die einzelnen Testaufgaben, T-Werte sowie die klassische Trennschärfe herangezogen (s. Kapitel 5.3.3).

9.1.1 DAS 1-DIMENSIONALE MODELL ZUR REPRÄSENTATIONSKOMPETENZ

Die statistischen Kennwerte der Testitems für die eindimensionale Skalierung können TABELLE 24 entnommen werden. Alle Items, mit einer Ausnahme, weisen für die eindimensionale Rasch-Skalierung eine gute Passung zum Modell auf, da sich die Fit-Maße innerhalb der festgelegten Grenzen befinden (s. Kapitel 5.3.3).

TABELLE 24: KENNWERTE DER EINDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 1-DIM-MODELL (N=548)

Skalierung			
Itemzahl	36		
Final Deviance	5250,83951		
Rel_i	0,94		
Varianz	0,47		
Rel_{EAP/PV}	0,42		
	S	Max.	Min.
wMNSQ		1,32	0,92
T-Werte		0,80	-0,80
r_{it}		0,02	0,56
I	3,95	2,70	-1,25
P	6,27	3,23	-3,04

Die für die eindimensionale Skalierung in der Hauptstudie ermittelten wMNSQ- ($0,92 \leq wMNSQ \leq 1,32$) und T-Werte ($-0,8 \leq T \leq 0,8$) zeigen, dass sich die Passung der Testitems etwas verbessert hat oder gleich geblieben ist (vgl. TABELLE 13, TABELLE 14). Dies trifft ebenfalls für die klassische Trennschärfe zu, die für alle Items oberhalb von 0,25 liegt. Eine Ausnahme stellt das Item See09 dar, welches mit einer Trennschärfe von $r_{it} = 0,02$ deutlich unter dem akzeptablen Grenzbereich bleibt. In ABBILDUNG 29 ist die Person-Item-Map der fallzentrierten Skalierung für das eindimensionale Repräsentationsmodell der untersuchten Testaufgaben bei einer Stichprobe von N=548 Studierenden dargestellt.

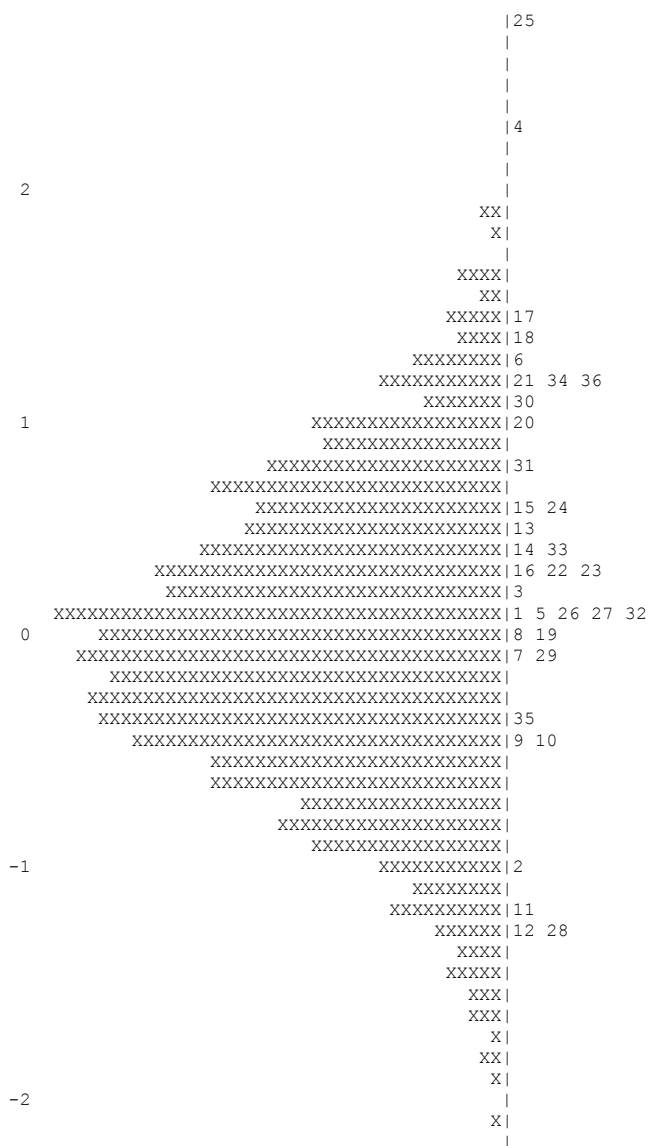


ABBILDUNG 29: EINDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG ZUR INTEGRATION VON MER (EIN X REPRÄSENTIERT 0,8 STUDIERENDE)

In der Wright-Map sind auf der linken Seite die Personenparameter verteilt, wobei jedes Kreuz 0,8 Fälle repräsentiert. Die Spanne der Itemschwierigkeit beträgt 3,95 Logits (Max.=2,70, Min.=-1,25) und liegt innerhalb des nach Rauch & Hartig (2008) üblichen Wertebereichs von -3 bis +3 Logits. Die Spanne der Personenfähigkeiten beträgt 6,95 Logits (Max.=3,23, Min.=-3,04). Damit ist die Stichprobe der Hauptstudie immer noch sehr heterogen, obwohl die Varianz geringer geworden ist. Sowohl im oberen als auch im unteren Bereich fehlen Testaufgaben, die das entsprechende Leistungsspektrum abdecken. Die Items sind für die leistungsstarken Probanden zum Teil nicht schwer genug oder für die leistungsschwachen Probanden zu schwer zu lösen. Die mittlere Itemschwierigkeit beträgt 0,38 Logits ($SD=0,87$), die mittlere Personenfähigkeit beträgt -0,36 Logits ($SD=1,05$). Die Werte zeigen, ergänzt durch die graphische Auswertung in der Wright Map, dass der Test insgesamt etwas zu schwer für die

untersuchte Stichprobe ist. Mit einem Wert von $Rel_I = 0,94$ hat sich die Item Separation Reliabilität aufgrund des besseren Verhältnisses zwischen Stichprobengröße und Antworthäufigkeit verbessert. Dies deutet auf eine hohe Genauigkeit der geschätzten Itemschwierigkeiten hin. Die EAP/PV Person Separation Reliabilität liegt mit einem Wert von $Rel_{EAP/PV} = 0,42$ immer noch im moderaten Bereich und konnte sich nicht verbessern. Ein Grund kann die geringer gewordene Varianz trotz größerer Stichprobe sein.

9.1.2 DAS 3-DIMENSIONALE MODELL ZUR MER-INTEGRATION

ABBILDUNG 30 zeigt die Person-Item-Map der fallzentrierten Rasch-Skalierung des dreidimensionalen Rasch-Modells zur *Integration von MER* (MERI, MERII, MERIII).

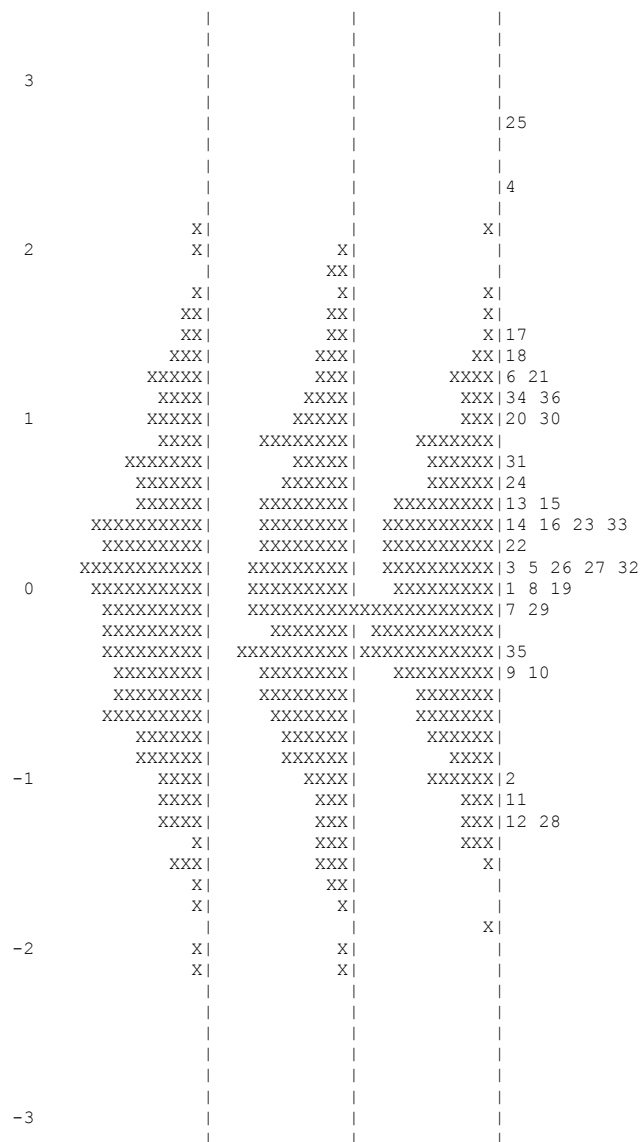


ABBILDUNG 30: DREIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 3-DIM-MODELL DER MER-INTEGRATION (EIN X STEHT FÜR 3,2 STUDIERENDE)

Die Testaufgaben verteilen sich auf die drei Dimensionen zur *Integration von multiplen externen Repräsentationen (MER)* und werden in *MERI (Informationsentnahmenähe Integration)*, *MERII (Konstruktionsnahe Integration)* und *MERIII (Translation und Transformation aus mehr als zwei MER)* differenziert (s. Kapitel 3.3). Die empirisch ermittelten Kennwerte werden für die drei Dimensionen getrennt angegeben und sind in TABELLE 25 dargestellt. Demzufolge passen die Items gut zum Rasch-Modell und genügen den Qualitätskriterien der probabilistischen Testtheorie (*MERI*: $0,95 \leq wMNSQ \leq 1,04$, $0,7 \leq t \leq 0,4$, $r_{it} > 0,25$; *MERII*: $0,91 \leq wMNSQ \leq 1,08$, $-1,0 \leq t \leq 1,0$, $r_{it} > 0,25$; *MERIII*: $0,96 \leq wMNSQ \leq 1,05$, $-0,7 \leq t \leq 0,7$, $r_{it} \geq 0,25$ mit Ausnahme von Item See09).

TABELLE 25: KENNWERTE DER DREIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 3-DIM MODELL DER MER-INTEGRATION

Skalierung									
Itemzahl	36								
Final Deviance	5242,98163								
Rel_i	0,94								
	MERI			MERII			MERIII		
Varianz	0,70			0,82			0,55		
Rel_{EAP/PV}	0,33			0,37			0,37		
	S	Max.	Min.	S	Max.	Min.	S	Max.	Min.
wMNSQ		1,04	0,95		1,08	0,91		1,05	0,96
T-Werte		0,40	-0,70		1,10	-1,00		0,70	-0,70
r_{it}		0,54	0,27		0,54	0,26		0,56	0,02
I	3,64	2,35	-1,29	1,60	1,54	0,06	4,04	2,74	-1,30
P	6,30	3,46	-2,84	5,07	3,17	-1,90	6,34	3,86	-2,48

Die Itemschwierigkeiten konnten mit einer Item Separation Reliabilität von $Rel_i = 0,94$ genau geschätzt werden. Alle Testaufgaben liegen in ihrer Itemschwierigkeit im üblichen Wertebereich von -3 bis +3 Logits (vgl. Rauch & Hartig, 2008). Das Item See09 als besonders schweres Item wurde zugunsten einer hinreichenden Inhaltsvalidität zur Abdeckung der *MERIII-Integration* als Prozess aus Informationsentnahme und Konstruktion im Itempool belassen (vgl. Bühner, 2006). Auf allen Dimensionen übersteigt die maximale Personenfähigkeit die

maximale Itemschwierigkeit. Die Items sind somit nicht in der Lage das gesamte Fähigkeitsspektrum der vorliegenden Stichprobe abzudecken. Auf allen drei Dimensionen fehlen schwierige Items, während im unteren Bereich leichte Items fehlen. Diese Heterogenität zeigt auch der Vergleich der Mittelwerte für die Dimension *MERI* ($M_{\text{Personenfähigkeit}}=0,01$, $SD_{\text{Personenfähigkeit}}=1,29$, $M_{\text{Itemschwierigkeit}}=-0,03$, $SD_{\text{Itemschwierigkeit}}=1,03$), und Dimension *MERII* ($M_{\text{Personenfähigkeit}}=0,22$, $SD_{\text{Personenfähigkeit}}=1,20$, $M_{\text{Itemschwierigkeit}}=0,71$, $SD_{\text{Itemschwierigkeit}}=0,47$) sowie der Dimension *MERIII* ($M_{\text{Personenfähigkeit}}=0,14$, $SD_{\text{Personenfähigkeit}}=1,21$, $M_{\text{Itemschwierigkeit}}=0,51$, $SD_{\text{Itemschwierigkeit}}=0,98$). Die EAP/PV Person Separation Reliabilität liegt für die Dimension *MERI* bei 0,33, für *MERII* bei 0,37 und für *MERIII* bei 0,37. Die gesunkenen Reliabilitäten können u.a. auf die geringe Verteilung der Testaufgaben auf die einzelnen Dimensionen zurückgeführt werden und sind Teil der Diskussion in Kapitel 10.2.

9.1.3 DAS 3-DIMENSIONALE KONTEXTMODELL ZUR MER-INTEGRATION

Die Person-Item-Map der fallzentrierten Skalierung für das dreidimensionale Kontextmodell (Ökologie, Stoffwechsel, Genetik) ist in ABBILDUNG 31 dargestellt. In dieser Skalierung wurden entsprechend den inhaltlichen und methodischen Annahmen (s. Kapitel 7.2 und Kapitel 8.1.3) die 36 Testaufgaben unterschiedlich auf die drei Aufgabenkontexte aufgeteilt. Jeweils neun Items für die Kontexte Ökologie und Genetik und 18 Items für den Kontext Stoffwechsel. Die doppelte Anzahl ergibt aus der Berücksichtigung der zwei Repräsentationstypen Diagramm und Schemazeichnung.

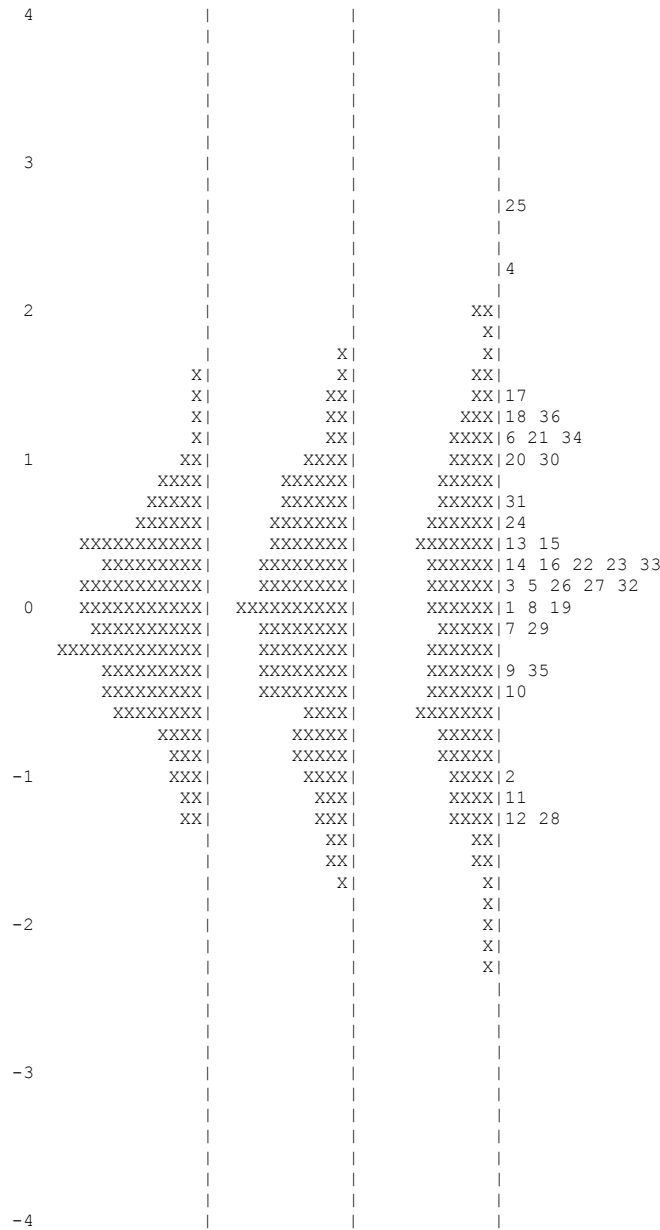


ABBILDUNG 31: DREIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 3-DIM-BETWEEN-MODELL DER AUFGABENKONTEXTE (EIN X STEHT FÜR 4,3 STUDIERENDE)

Die ermittelten Kennwerte zeigen, dass die Items der Dimension *Ökologie* ($0,97 \leq wMNSQ \leq 1,07$, $-0,4 \leq t \leq 1,1$, $r_{it} > 0,35$, Ausnahme Item See09), *Stoffwechsel* ($0,95 \leq wMNSQ \leq 1,07$, $-0,7 \leq t \leq 0,7$, $r_{it} > 0,25$) und *Genetik* ($0,96 \leq wMNSQ \leq 1,06$, $-0,7 \leq t \leq 0,8$, $r_{it} > 0,25$) den beschriebenen Qualitätskriterien der probabilistischen Testtheorie genügen und insgesamt gut zum Rasch-Modell passen (vgl. TABELLE 26).

TABELLE 26: KENNWERTE DER DREIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 3DIM-BETWEEN-MODELL DER AUFGABENKONTEXTE ÖKOLOGIE, STOFFWECHSEL, GENETIK

Skalierung									
Itemzahl	36								
Final Deviance	5241,35686								
Rel_I	0,94								
	Ökologie			Stoffwechsel			Genetik		
Varianz	0,37			0,64			1,09		
Rel_{EAP/PV}	0,35			0,38			0,33		
	S	Max.	Min.	S	Max.	Min.	S	Max.	Min.
wMNSQ		1,07	0,97		1,07	0,95		1,06	0,96
T-Werte		1,10	-0,40		0,70	-0,70		0,80	-0,70
r_{it}		0,54	0,02		0,56	0,26		0,54	0,27
I	3,64	2,66	-0,98	3,58	2,31	-1,27	2,67	1,29	-1,38
P	6,07	3,78	-2,29	6,24	3,47	-2,77	5,96	3,14	-2,82

Die Item Separation Reliabilität zeigt einen Wert von $Rel_I = 0,94$ und weist für alle Modellierungen auf eine hohe Genauigkeit der geschätzten Aufgabenschwierigkeiten hin. Die EAP/PV Person Separation Reliabilität liegt für alle drei Dimensionen in einem noch moderaten Bereich (Ökologie: $Rel_{EAP/PV} = 0,35$, Stoffwechsel: $Rel_{EAP/PV} = 0,38$, Genetik: $Rel_{EAP/PV} = 0,33$). Analog zu den anderen mehrdimensionalen Modellen lassen sich nur wenig weitere Unterschiede feststellen. Im Hinblick auf die Varianz zeigen sich sehr heterogene Ergebnisse. Innerhalb der Genetikaufgaben ist die Varianz mit einem Wert von $v = 1,09$ deutlich höher als für die Aufgaben aus dem Bereich der Stoffwechselphysiologie ($v = 0,64$) und der Ökologie ($v = 0,37$). Dies bedeutet, dass den Testitems insgesamt eine eingeschränkte Variation an Itemantworten zugrunde liegt. Dies gilt insbesondere für den Kontext Ökologie, während für den Kontext Genetik zwischen den Fähigkeiten relativ gut unterschieden werden kann. Aufgrund des Zusammenhangs zwischen Varianz und den EAP/PV Person Reliabilitäten der Skalen sowie zur Anzahl der zugehörigen Items (Rost, 2004; vgl. Wellnitz, 2012, S. 113), sind diese Befunde näher zu diskutieren (s. Kapitel 10). Für alle Dimensionen übersteigt die maximale Personen-

fähigkeit die maximale Itemschwierigkeit. Die Abdeckung mit Testitems gelingt in den Aufgabenkontexten Ökologie und Stoffwechsel im oberen Bereich noch relativ gut (etwa 1,1 Logits). Im unteren Schwierigkeitsbereich fehlten hingegen Items (vgl. ABBILDUNG 31).

9.1.4 DAS 2-DIMENSIONALE REPRÄSENTATIONSMODELL ZUR MER-INTEGRATION

Für die Verteilung der latenten Konstrukte im zweidimensionalen Modell werden die Fähigkeiten zwischen Diagrammkompetenz und Schemakompetenz unterschieden. Die Person-Item-Map der fallzentrierten Rasch-Skalierung zur zweidimensionalen Betrachtung des Repräsentationsmodells (Diagramm, Schema) ist in ABBILDUNG 32 dargestellt.

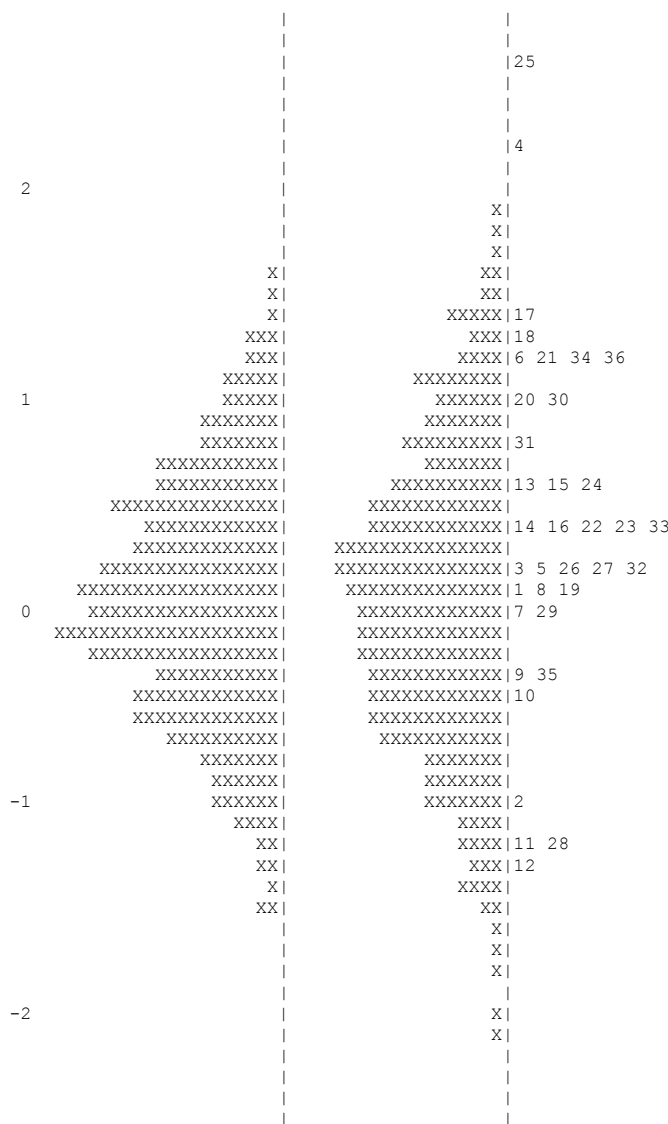


ABBILDUNG 32: ZWEIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 2-DIM-BETWEEN-MODELL REPRÄSENTATIONSTYP DIAGRAMM UND SCHEMA (EIN X STEHT FÜR 2,1 STUDIERENDE)

Die Kennwerte werden für die zwei Dimensionen getrennt angegeben und finden sich in TABELLE 27. Die ermittelten Kennwerte zeigen, dass die Items der Dimensionen *Diagrammkompetenz* ($0,96 \leq wMNSQ \leq 1,08$, $-0,7 \leq t \leq 1,2$, $r_{it} > 0,25$, Ausnahme Item See09) und *Schema-kompetenz* ($0,93 \leq wMNSQ \leq 1,10$, $-0,9 \leq t \leq 1,0$, $r_{it} > 0,25$) gut zum Rasch-Modell passen.

TABELLE 27: KENNWERTE DER ZWEIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 2DIM-BETWEEN-MODELL DER REPRÄSENTATIONSTYPEN DIAGRAMM UND SCHEMA

		Skalierung					
Itemzahl	36						
Final Deviance	5249,52514						
Rel_I	0,94						
		Diagramm			Schema		
Varianz		0,42			0,62		
Rel_{EAP/PV}		0,38			0,40		
		S	Max.	Min.	S	Max.	Min.
wMNSQ			1,08	0,96		1,10	0,93
T-Werte			1,20	-0,70		1,00	-0,90
r_{it}			0,54	0,02		0,56	0,26
I	3,91	2,68	-1,23		2,49	1,20	-1,29
P	6,50	3,95	-2,55		6,13	3,26	-2,87

Die Item Separation Reliabilität liegt bei $Rel_I = 0,94$. Die EAP/PV Person Separation Reliabilität liegt für beide Dimension in einem moderaten Bereich (Diagramm: $Rel_{EAP/PV} = 0,38$, Schema: $Rel_{EAP/PV} = 0,40$) und unterscheidet sich kaum von den Reliabilitäten der anderen Dimensionsmodellierungen. In beiden Dimensionen übersteigt die maximale Personenfähigkeit die maximale Itemschwierigkeit. Während für die Diagrammkompetenz das Leistungsspektrum im oberen Bereich noch mit etwa 1,0-1,5 Logits Unterschied moderat abgedeckt werden kann, fehlen vor allem Items im unteren Schwierigkeitsbereich. Für die Schemakompetenz ist die Differenz in der Itemschwierigkeit noch größer und es fehlen sowohl im unteren als auch im oberen Bereich Testaufgaben. Dies zeigt auch der Vergleich der Mittelwerte für die Dimensionen *Diagramm* ($M_{Personenfähigkeit}=0,05$, $SD_{Personenfähigkeit}=1,20$; $M_{Itemschwierigkeit}=0,57$ $SD_{Itemschwierigkeit}=0,97$) und *Schema* ($M_{Personenfähigkeit}=0,00$, $SD_{Personenfähigkeit}=1,21$; $M_{Itemschwierigkeit}=0,20$ $SD_{Itemschwierigkeit}=0,74$). Somit fehlen insgesamt vor allem leichte Testaufgaben (vgl. ABBILDUNG 32).

9.1.5 DAS 2-DIMENSIONALE MODELL ZUR MER-INTEGRATION

Unter Rückbezug auf die beschriebenen Messmodelle in Kapitel 8.1.3 Messmodelle der Untersuchung wird im Folgenden die fallzentrierte Rasch-Skalierung für die Annahme einer 2-Dimensionalität der *MER-Integration* vorgestellt. Für die Skalierung werden alle zwölf Testaufgaben der Dimension MERIII entweder der Dimension MERI oder der Dimension MERII zugeordnet. Die Zuordnung erfolgt nach dem Ladungsmuster der Between-Item-Mehrdimensionalität und richtet sich nach dem Output Text (MERI) oder Bild (MERII) im Integrationsprozess.

Die Person-Item-Map des zweidimensionalen Rasch-Modells zur *MER-Integration* ist in ABILDUNG 33 dargestellt. Die empirischen Kennwerte finden sich in TABELLE 28.

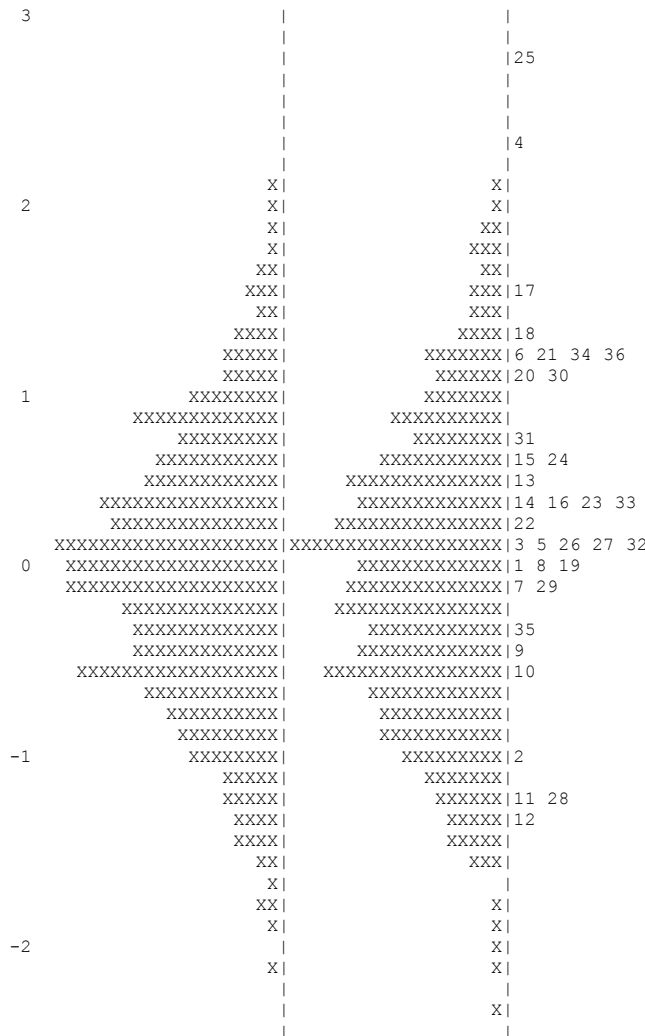


ABBILDUNG 33: ZWEIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 2-DIM-BETWEEN-MODELL DER MER-INTEGRATION (EIN X STEHT FÜR 1,9 STUDIERENDE)

Die Itemkennwerte zeigen, dass die Testaufgaben der Dimension *MERI* und *MERII* gut zum Rasch-Modell passen (*MERI*: $0,94 \leq wMNSQ \leq 1,08$, $-0,6 \leq t \leq 1,0$, $r_{it} > 0,25$; *MERII*: $0,91 \leq wMNSQ \leq 1,08$, $-0,7 \leq t \leq 1,0$, $r_{it} > 0,25$ mit Ausnahme von Item See09).

TABELLE 28: KENNWERTE DER ZWEIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 2-DIM-BETWEEN-MODELL DER MER-INTEGRATION

Skalierung						
Itemzahl	36					
Final Deviance	5242,50984					
Rel_I	0,94					
	MERI			MERII		
Varianz	0,57			0,73		
Rel_{EAP/PV}	0,37			0,36		
	S	Max.	Min.	S	Max.	Min.
wMNSQ	1,08 0,94			1,08 0,91		
T-Werte	1,00 -0,60			1,00 -0,70		
r_{it}	0,54 0,27			0,56 0,02		
I	3,56	2,29	-1,27	3,10	2,79	-0,31
P	6,43	3,53	-2,90	5,95	3,96	-1,99

Die Item Separation Reliabilität ($Rel_I = 0,94$) sowie die EAP/PV Person Separation Reliabilität ($MERI\ Rel_{EAP/PV} = 0,37$; $MERII\ Rel_{EAP/PV} = 0,36$) zeigen im Vergleich zur dreidimensionalen Rasch-Skalierung keine besseren Werte (vgl. TABELLE 25). Die Itemschwierigkeiten liegen analog zu den anderen Rasch-Skalierungen innerhalb des üblichen Wertebereichs von -3 bis +3 Logits (Rauch & Hartig, 2008). Die Kennwerte zeigen, dass das Leistungsspektrum der MER-Integration der Studierendenfähigkeiten von 6,27 Logits bei einer Spanne von 3,93 Logits in der Itemschwierigkeit relativ gut abgedeckt wird. Die Itemparameter der mehrdimensionalen Skalierungen weisen auch für die zweidimensionale Skalierung zum Teil größere, zum Teil kleinere Schwierigkeitsspannen auf. In beiden Dimensionen übersteigt die maximale Personenfähigkeit die maximale Itemschwierigkeit und umgekehrt. Entsprechend fehlen leichte und schwierige Items. Dies wird durch den Vergleich der Mittelwerte für die Dimension *MERI* ($M_{Personenfähigkeit}=0,02$, $SD_{Personenfähigkeit}=1,21$; $M_{Itemschwierigkeit}=0,05$ $SD_{Itemschwierigkeit}=0,88$) und Dimension *MERII* ($M_{Personenfähigkeit}=0,17$, $SD_{Personenfähigkeit}=1,20$; $M_{Itemschwierigkeit}=0,83$, $SD_{Itemschwierigkeit}=0,47$) noch deutlicher. Während die Items der *MERI*-Dimension relativ gut für die Stichprobe passen, sind die Items der *MERII*-Dimension insgesamt zu schwer für die getesteten Personen.

9.1.6 DAS 6-DIMENSIONALE WITHIN-MODELL ZUR MER-INTEGRATION

Abschließend wird die Within-Modellierung für die sechsdimensionale, fallzentrierte Rasch-Skalierung für die Dimensionen der *MER-Integration* (MERI, MERII, MERIII) sowie für die drei *Aufgabenkontexte* (Ökologie, Stoffwechsel und Genetik) vorgestellt. In dieser Skalierung wird berücksichtigt, dass die Testaufgaben jeweils auf die latente Dimension *Informationsentnahmenähe Integration (MERI)* oder *Konstruktionsnahe Integration (MERII)* oder *Translation und Transformation aus mehr Repräsentationen (MERIII)* laden und gleichzeitig einem der drei Aufgabenkontexte zugeordnet werden können. Als Ergebnis der fallzentrierten Rasch-Skalierung kann festgehalten werden, dass bei einer Itemzahl von 36 die Final Deviance 5230,28676 beträgt. Das ist der insgesamt niedrigste Wert für alle durchgeführten Skalierungen zur Modellprüfung. Die Item Separation Reliabilität bleibt mit einem Wert von $Rel_I = 0,94$ auf einem hohen Niveau. Für die Dimensionen MERI beträgt die Varianz $\nu = 0,64$, für die Dimension MERII $\nu = 0,70$ und für die Dimension MERIII $\nu = 0,40$. Die Varianzen der Aufgabenkontexte zeigen Werte von $\nu = 0,25$ für die Dimension Ökologie, $\nu = 0,41$ für die Dimension Stoffwechsel und $\nu = 0,71$ für die Dimension Genetik. Für diese Rasch-Skalierung mit Within-Item-Mehrdimensionalität konnten keine EAP/PV Person Separation Reliabilitäten ausgegeben werden. Die wMNSQ-Werte ($0,95 \leq wMNSQ \leq 1,06$) sowie die T-Werte ($-0,7 \leq t \leq 0,9$) sind hingegen zufriedenstellend. Die Itemschwierigkeit zeigt eine Spanne von insgesamt 2,08 Logits (Min. = -0.71; Max._t=1.37).

9.1.7 ZUSAMMENFASSUNG DER MODELLPASSUNG

Zusammenfassend lässt sich für die Prüfung und Validierung der empirischen Daten auf ihre Modellgeltung festgehalten, dass die empirischen Kennwerte der 36 Testaufgaben den üblichen Qualitätskriterien für probabilistische Berechnungen genügen und durch die geschätzten Rasch-Modelle passend beschrieben werden (vgl. Bond & Fox, 2007; Rost, 2004; Wu & Adams, 2007). Zugunsten einer hinreichenden Inhaltsvalidität (vgl. Bühner, 2006) wird das gemäß den Kennwerten kritische Item See09 im Itempool belassen. Damit kann davon ausgegangen werden, dass die Testitems zur Erfassung von Kompetenzen geeignet sind und der Test in allen Dimensionen und Subdimensionen das gesamte Fähigkeitsspektrum der zugrundeliegenden Stichprobe erfasst (vgl. Bühner, 2006).

Zusätzlich zur Modellkontrolle wurde überprüft, ob die Item- und Personenparameter jeweils normalverteilt sind. Der Shapiro-Wilk-Test ist für die Itemparameter nicht signifikant ($p > 0,05$). Entsprechend kann von einer Normalverteilung der Testaufgaben ausgegangen werden. Die Personenparameter weisen hingegen keine Normalverteilung in den Daten auf. Aus

diesem Grund werden für Anschlussanalysen nicht-parametrische Testverfahren angewendet (s. Kapitel 5.3.4). Aufgrund der nur ausreichenden Reliabilitäten für die Skalierungen aller einzelnen Modellierungen sollten die Daten vorsichtig interpretiert werden.

9.2 MODELLVERGLEICH

Auf Grundlage des Kompetenzmodells (s. Kapitel 3.3) sowie der vorgenommenen Aufgabekonstruktion (s. Kapitel 4.2.2 und Kapitel 7.2) ergeben sich sechs mögliche Modellvarianten: ein eindimensionales Gesamtmodell zur Repräsentationskompetenz, ein dreidimensionales Modell der MER-Integration (MERI, MERII, MERIII), ein dreidimensionales Kontextmodell (Ökologie, Stoffwechsel, Genetik), ein zweidimensionales Repräsentationsmodell (Diagramm, Schema), ein zweidimensionales Modell der MER-Integration (MERI, MERII) als Alternativmodell sowie ein sechsdimensionales Within-Modell. Für die Modellprüfung wurde eine itemzentrierte Rasch-Skalierung durchgeführt. Die itemzentrierte Rasch-Skalierung dient zur Ermittlung der Parameterzahl und der Final Deviance (s. Kapitel 8.1.2). Diese werden für die Berechnung der informationstheoretischen Maße AIC, BIC und CAIC herangezogen, die eine Entscheidung über die passendste Modellvariante erlauben (s. Kapitel 8.2.2).

Die Kennwerte der itemzentrierten Skalierungen für die einzelnen Modelle sind in TABELLE 29 zusammenfassend dargestellt. Erwartungsgemäß haben sich die Item Separation Reliabilitäten, Varianzen, EAP/PV Person Separation Reliabilitäten und die wMNSQ- und t-Werte sowie die klassische Trennschärfe im Vergleich zur fallzentrierten Skalierung nur marginal verändert.

TABELLE 29: KENNWERTE DER ITEMZENTRIERTEN RASCH-SKALIERUNGEN FÜR DIE EIN- UND MEHRDIMENSIONALEN MODELLIERUNGEN

	1D		3D MER				3D CON			
Final Deviance	5250,84016		5242,98163				5241,35677			
Parameterzahl	37		42				42			
Rel_i	0,97		0,97				0,97			
Varianz	0,47		0,70	0,82	0,55	0,37	0,64	1,09		
Rel_{EAP/PV}	0,42		0,33	0,37	0,37	0,35	0,38	0,33		
	2D TYP		2D MER		6D within MERCON					
Final Deviance	5247,91531		5242,50987		5230,28676					
Parameterzahl	39		39		57					
Rel_i	0,97		0,97		0,94					
Varianz	0,42	0,62	0,57	0,73	0,64	0,67	0,40	0,25	0,41	0,71
Rel_{EAP/PV}	0,38	0,40	0,37	0,36	-	-	-	-	-	-

Die Modellprüfungen finden auf der Ebene der geschätzten Personenkenwerte statt (Wu et al., 2007). Zunächst werden die Final Deviance der unterschiedlichen Modelle verglichen. Je kleiner der Wert, desto besser passt das Modell auf die Daten (Wu & Adams, 2007). Das sechsdimensionale Modell aus *MER-Integration* (MERI, MERII, MERIII) und den *Aufgabenkontexten* (Ökologie, Stoffwechsel, Genetik) mit Within-Ladungsmuster zeigt die niedrigste Final Deviance. Demzufolge liegt in diesem Modell die geringste Abweichung von den ermittelten Daten vor. Die Passung wird mit 57 Parametern erreicht. Die dreidimensionalen Modelle *MER-Integration* und *Aufgabenkontext* sowie die zweidimensionale Modellierung *MER-Integration*, jeweils mit Between-Item-Mehrdimensionalität, haben eine Final Deviance von 5242,98163 und 5241,35677 sowie von 5242,50987. Gleichzeitig kommen diese Modelle mit einer jeweils geringen Parameterzahl von 42 bzw. 39 aus. Auch die Final Deviance der zweidimensionalen Modellierung für den *Repräsentationstyp* ist höher (5247,91531) und hat die zweithöchste Final Deviance insgesamt. Für das eindimensionale Modell zur *Repräsentationskompetenz* gibt ConQuest die höchste Final Deviance aus (5250,84016). Die im Vergleich zu den anderen Modellen schlechtere Passung zu den Daten wird hingegen mit nur 37 Parametern erreicht. Vor diesem Hintergrund folgt der Vergleich der verschiedenen Modellvarianten mithilfe der informationstheoretischen Maße AIC; BIC und CAIC (vgl. TABELLE 30). Für den

Modellvergleich wird vorrangig der BIC herangezogen, da in diesem die Stichprobengröße, die Testlänge und die kleinen Häufigkeiten der beobachteten Antwortmuster berücksichtigt werden (vgl. Rost, 2004; Gollwitzer, 2008; Wellnitz, 2012).

TABELLE 30: MODELLPRÜFUNG NACH DEM AIC, BIC UND CAIC

Modell	Final Deviance	Parameter	AIC	BIC	CAIC
6D within MERCON	5230,28676	57	5344,28676	5386,39816	5443,39725
3D CON	5241,35677	42	5325,35677	5356,38555	5398,38555
2D MER	5242,50987	39	5320,50987	5349,32231	5388,32231
3D MER	5242,98163	42	5326,98163	5358,01041	5400,01041
2D TYP	5247,91531	39	5325,91531	5354,72775	5393,72775
1D	5250,84016	37	5324,84016	5352,17504	5389,17504

Setzt man die Final Deviance und die Parameterzahl unter Berücksichtigung der Stichprobengröße ins Verhältnis besitzt das Modell mit dem geringsten BIC die beste Passung zu den empirischen Daten. Daraus resultiert $BIC_{2DMER} < BIC_{1D} < BIC_{2DTYP} < BIC_{3DCON} < BIC_{3DMER} < BIC_{6Dwithin}$. Das zweidimensionale Modell der MER-Integration weist nach dem BIC im Vergleich zu allen anderen die größte Übereinstimmung mit den Daten auf. Unter Berücksichtigung des Kriteriums, dass die Differenz zwischen den informationstheoretischen Maßen zwei zu vergleichender Modelle zehn betragen sollte, um von einer besseren Modellpassung zu sprechen, zeigt sich, dass alle Modelle mit Ausnahme des 6D-within-Modells gleich gut zu den Daten passen. In diesem Zusammenhang ist festzuhalten, dass das 6D-within-Modell die insgesamt geringste Final Deviance zeigt (vgl. auch TABELLE 29). Wird die Parameterzahl berücksichtigt, resultiert ein insgesamt höherer BIC im Vergleich zu allen anderen Modellen. Wie Rost (2004) berichtet, können Modellgeltungstests schon ab zwei Dimensionen problematisch werden. Vor diesem Hintergrund sollte die Ablehnung dieses Modells mit Vorsicht getroffen werden. Aus den Befunden kann jedoch abgeleitet werden, dass sich die Kompetenzen zur MER-Integration in mehrere Teilkompetenzen differenzieren lassen. Unter Rückbezug auf das Sparsamkeitskriterium fällt der Modellvergleich abschließend zugunsten des 2D-MER-Modells aus, da es gemäß der informationstheoretischen Maße mit einer

Für eine weitere Präzisierung der Modellprüfung werden die latenten Korrelationen innerhalb der mehrdimensionalen Kompetenzkonstrukte herangezogen. Diese korrelationsstatistischen Analysen dienen einer weiteren Absicherung, ob sich die Subdimensionen als separate Teilkompetenzen unterscheiden lassen (Wellnitz, 2012). Falls latente Korrelationen unter $r = 0,90$ liegen, kann von einem mehrdimensionalen Konstrukt ausgegangen werden (Bond & Fox, 2007; Roeling, 2016). Für das zweidimensionale Modell liegt die latente Korrelation zwischen den Dimensionen MERI und MERII im mittleren Bereich bei $r = 0,57$ (vgl. TABELLE 31). Daneben werden auch die Korrelationen zwischen den Personenparametern angegeben, um eine Überschätzung der latenten Korrelationen auszuschließen (vgl. Schmiemann, 2010). Die Korrelation zwischen den Personenparametern beträgt $r = 0,37$ und kann als gering eingestuft werden (vgl. Zöfel, 2003). Die tatsächliche Korrelation ist demnach zwischen diesen beiden Werten anzusiedeln (Schmiemann, 2010).

TABELLE 31: LATENTE KORRELATION (\diamond) ZWISCHEN DEN DIMENSIONEN MERI UND MERII (\diamond) BZW. KORRELATION DER PERSONENPARAMETER

Dimension	MERI	MERII
MERI	–	0,37
MERII	0,57 \diamond	–

Im Folgenden werden die korrelationsstatistischen Analysen auch für das dreidimensionale MER-Modell (MERI, MERII, MERIII), das dreidimensionale Kontextmodell (Ökologie, Stoffwechsel, Genetik) und das sechsdimensionale Within-Modell dargestellt.

In TABELLE 32 finden sich die Korrelationen für das 3D-MER-Modell. Demnach liegt eine geringe Korrelation zwischen den beiden Dimensionen MERI und MERII ($r = 0,48$) und eine mit $r = 0,68$ mittlere Korrelation zwischen den Dimensionen MERI und MERIII vor. Die Korrelation zwischen MERII und MERIII liegt im hohen Bereich ($r = 0,77$). Die Korrelationen zwischen den Personenparametern liegen ebenfalls im geringen bis mittleren Bereich.

TABELLE 32: LATENTE KORRELATIONEN (\diamond) ZWISCHEN DEN DIMENSIONEN MERI, MERII, MERIII BZW. KORRELATION DER PERSONENPARAMETER

Dimension	MERI	MERII	MERIII
MERI	–	0,37	0,42
MERII	0,48 \diamond	–	0,52
MERIII	0,68 \diamond	0,77 \diamond	–

Wie aus TABELLE 33 ersichtlich wird, sind die Leistungen der Studierenden in den verschiedenen Aufgabenkontexten durch mittlere bis hohe Korrelationen gekennzeichnet. So liegt zwischen dem Kontext Ökologie und Stoffwechsel eine latente Korrelation von $r = 0,8$ und damit hohe Korrelation vor. Die Korrelation zwischen Ökologie und Genetik liegt im mittleren Bereich ($r = 0,64$), ebenfalls die Korrelation zwischen dem Kontext Stoffwechsel und Genetik ($r = 0,51$). Unter Berücksichtigung der Korrelationen zwischen den Personenparametern lassen sich hingegen insgesamt eher geringe Korrelationen für das dreidimensionale Kontextmodell 3D-CON ableiten.

TABELLE 33: LATENTE KORRELATIONEN (\diamond) ZWISCHEN DEN DIMENSIONEN ÖKO, STW, GEN BZW. KORRELATION DER PERSONENPARAMETER

Dimension	ÖKO	STW	GEN
ÖKO	–	0,40	0,41
STW	0,82 \diamond	–	0,43
GEN	0,64 \diamond	0,51 \diamond	–

Für den Zusammenhang zwischen MER-Integration und den Aufgabenkontexten sowie vor dem Hintergrund der resultierenden Werte aus dem Vergleich der informationstheoretischen Maße, wird im Folgenden die korrelationsstatistische Analyse für das 6D-within-MERCON-Modell dargestellt. Die Kennwerte finden sich in TABELLE 34.

TABELLE 34: LATENTE KORRELATIONEN (\diamond) ZWISCHEN DEN DIMENSIONEN MERI, MERII, MERIII UND DEN AUFGABENKONTEXTEN (ÖKO, STW, GEN) FÜR DAS 6D-WITHIN-MODELL BZW. KORRELATION DER PERSONENPARAMETER

Dimension	MERI	MERII	MERIII	ÖKO	STW	GEN
MERI	–	0,21	0,26	0,00	-0,01	-0,03
MERII	0,32 \diamond	–	0,31	-0,15	-0,07	0,16
MERIII	0,51 \diamond	0,59 \diamond	–	-0,05	0,03	-0,08
ÖKO	0,01 \diamond	-0,36 \diamond	-0,16 \diamond	–	0,18	0,10
STW	-0,02 \diamond	-0,16 \diamond	0,07 \diamond	0,57 \diamond	–	0,10
GEN	-0,05 \diamond	0,23 \diamond	-0,14 \diamond	0,25 \diamond	0,19 \diamond	–

Die Korrelationen zwischen den Dimensionen der MER-Integration sowie zwischen der Dimension Aufgabenkontext decken sich weitgehend mit den Befunden aus den jeweils ge-

trennt berichteten Korrelationsanalysen (vgl. TABELLE 32 und TABELLE 33). So liegt die Korrelation zwischen MERI und MERII im niedrigen Bereich ($r = 0,32$) bzw. zwischen MERI und MERIII ($r = 0,51$) sowie MERII und MERIII ($r = 0,59$) im mittleren Bereich. Die Korrelationen zwischen den Aufgabenkontexten sind im Vergleich weiter gesunken. Während zwischen dem Kontext Ökologie und Stoffwechsel mit $r = 0,57$ eine mittlere Korrelation vorliegt, zeigt sich zwischen dem Kontext Ökologie und Genetik eine geringe ($r = 0,25$) bzw. zwischen Stoffwechsel und Genetik eine nur sehr geringe Korrelation von $r = 0,19$. Im Hinblick auf die Korrelationen zwischen MER-Integration und Aufgabenkontext kann festgehalten werden, dass sich diese im sehr geringen bis geringen Bereich befinden, was auf keine großen Gemeinsamkeiten hindeutet (vgl. TABELLE 34). Eine differenzierte Betrachtung der Korrelationen zwischen den Personenparameter wird aufgrund der einheitlich vorliegenden geringen bis sehr geringen Korrelationen ebenfalls nicht vorgenommen.

Hinsichtlich der Modellprüfung und des anschließenden Modellvergleichs kann festgehalten werden, dass die informationstheoretischen Maße und die geringen bis mittleren Interkorrelationen der Hauptstudie zwar auf Gemeinsamkeiten hindeuten, die ermittelten Daten dennoch auf eine Differenzierung in den Dimensionen und Subdimensionen verweisen.

9.3 SCHWIERIGKEITSUNTERSCHIEDE UND VORHERSAGE DER SCHWIERIGKEIT

Im Folgenden wird der Einfluss verschiedener Aufgabenmerkmale auf die Itemschwierigkeit untersucht. Für die Analyse werden die in der eindimensionalen, fallzentrierten Rasch-Skalierung ermittelten Itemschwierigkeiten genutzt. Mithilfe der ein- und mehrfaktoriellen Varianzanalyse wird geklärt, ob sich die mittlere Aufgabenschwierigkeit der im Kompetenzstrukturmodell definierten Komponenten *MER-Integration*, *Aufgabenkontext* und *Repräsentationstyp* auch in den Ergebnissen der Hauptstudie statistisch voneinander unterscheiden. Als unabhängige Variable werden jeweils erneut die Kategorien „MERI oder MERII“ sowie „Ökologie, Stoffwechsel oder Genetik“ und „Diagramm oder Schema“ als Gruppierungsvariablen definiert und in Bezug auf die abhängige Variable „Lösungswahrscheinlichkeit“ untersucht. Im Hinblick auf die Vorhersage der Itemschwierigkeit werden die Kategorien dahingehend untersucht, inwiefern diese die Schwierigkeit der Testaufgaben modellkonform vorhersagen. Um den Einfluss zu prüfen, werden multiple Regressionsanalysen durchgeführt (vgl. Hartig, 2007).

Für einen adäquaten Einsatz der Testverfahren und als Voraussetzung der Residuen für lineare Regressionsmodelle wurde zu Beginn überprüft, ob die Itemschwierigkeiten im eindimensionalen Modell normalverteilt sind (vgl. Field, 2009). Die Prüfung mit dem Shapiro-Wilk-Test

ergab keine signifikanten Abweichungen der Itemschwierigkeiten von einer Normalverteilung ($p > 0,05$). Die Voraussetzungen zur Durchführung der Regressionsanalyse wie Durbin-Watson-Koeffizient und Kollinearität via Toleranz und VIF (Variance Inflation Factor) wurden überprüft und können als akzeptabel angenommen werden. Die Normalverteilung der Residuen als Voraussetzung (vgl. Field, 2009) wird durch die grafische Testung bestätigt. Damit können die Regressionsmodelle als Ganzes und die einzelnen Regressionskoeffizienten auf ihre statistische Signifikanz überprüft werden.

9.3.1 MER-INTEGRATION

Um die Itemschwierigkeiten hinsichtlich signifikanter Unterschiede zu prüfen, wird zunächst eine einfaktorielle Varianzanalyse (ANOVA) für die drei MER-Integrationskompetenzen MERI, MERII und MERIII durchgeführt. Im Anschluss daran wird aufgrund des favorisierten zweidimensionalen MER-Integrationsmodells (MERI, MERII) ein t-Test für unabhängige Stichproben eingesetzt. Die Nullhypothese, dass die Items aller MER-Integrationskompetenzen gleich schwer sind, wird für das dreidimensionale Integrationsmodell mit einem Effekt von $F(2,33) = 2,280$ ($p > 0,05$, $\eta_p^2 = 0,12$) angenommen und für die zweidimensionale Modellierung (MERI, MERII) mit einem signifikanten Effekt von $t(34) = -2,775$ ($p < 0,01$, $d = 0,74$) verworfen. Die mittlere Itemschwierigkeit beträgt $M_\sigma = 0,38$ Logits ($SD=0,15$). Betrachtet man den gesamten Test, so liegt die Schwierigkeit der Items der *Informationsentnahmenahen Integration (MERI)* um 0,64 Logits unter der Schwierigkeit von Aufgaben zur *Konstruktionsnahen Integration (MERII)* und 0,47 Logits unter der Schwierigkeit von Aufgaben zur *Translation und Transformation aus mehr als zwei Repräsentationen (MERIII)*.

In ABBILDUNG 34 und ABBILDUNG 35 sind jeweils die Zusammenhänge zwischen den Aufgabenmerkmalen der Dimension MER-Integration und der Itemschwierigkeit dargestellt. Die Schwierigkeit der Items der MERI- und der MERIII-Integration zeigt eine im Vergleich etwas größere Box und längere Antennen, was auf eine stärkere Streuung dieser Items hindeutet. Die Kreise mit den Ziffern 24 und eins sind Ausreißer, d.h. Werte, deren Abstand vom 75%-Perzentil nach oben zwischen dem 1,5- bis 3-fachen der Boxhöhe liegt. Es handelt sich dabei um die beiden Items See09 und Anp03, die insgesamt die schwierigsten Werte im vorliegenden Test aufweisen. Dabei sind Aufgaben, die eine konstruktionsnahe Integration erfordern für die untersuchte Stichprobe am schwierigsten zu lösen ($M_{MERII}=0,67$, $SD=0,44$).

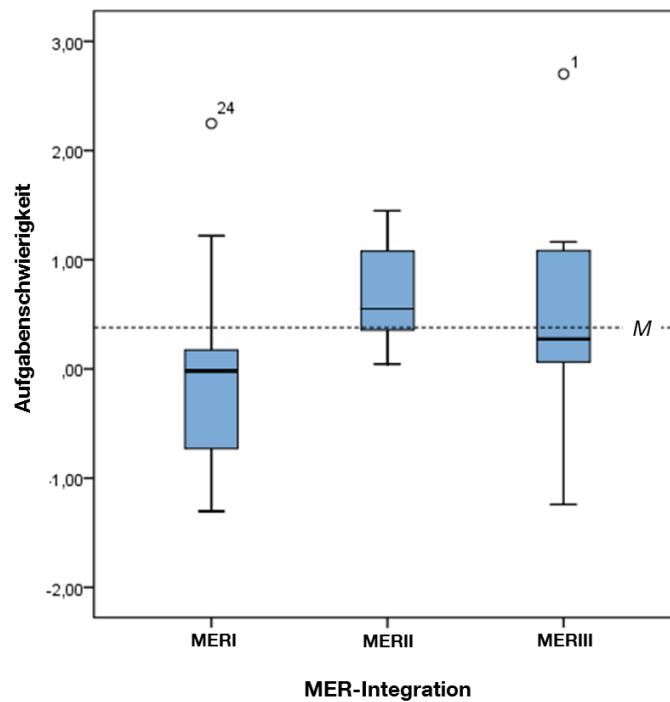


ABBILDUNG 34: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH KOMPONENTEN DER MER-INTEGRATION (DREIDIMENSIONAL, $N=36$ ITEMS, M = MITTELWERT)

Im zweidimensionalen MER-Integrationsmodell wird der Unterschied zwischen Informationsentnahmenaher Integration (MERI) und Konstruktionsnaher Integration (MERII) größer. Der Effekt ist statistisch bedeutsam. Dabei liegt die mittlere Itemschwierigkeit von MERI-Aufgaben bei $M_{MERI}=0,05$ ($SD=0,87$) und von MERII-Aufgaben bei $M_{MERII}=0,79$ ($SD=0,71$).

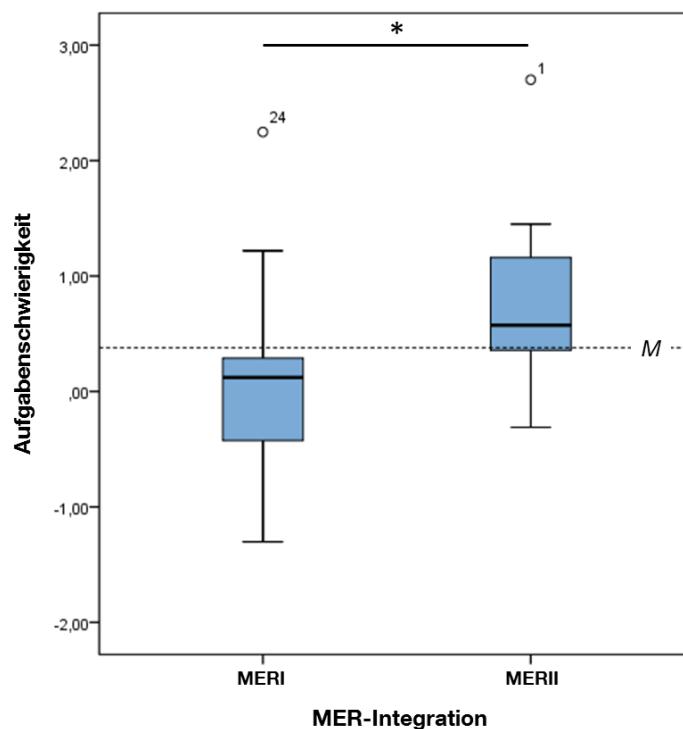


ABBILDUNG 35: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH KOMPONENTEN DER MER-INTEGRATION (ZWEIDIMENSIONAL, $N=36$ ITEMS, M = MITTELWERT)

Um zu klären, inwiefern die Itemschwierigkeit durch die MER-Integration vorhergesagt werden kann, wird eine einfache lineare Regression durchgeführt. Dadurch kann beispielsweise vorhergesagt werden, um wie viele Logits ein Item mit einem schwierigkeitsbestimmenden Merkmal, z.B. die MERII-Integration, schwieriger ist als ein Item ohne dieses Merkmal (vgl. Ziepprecht, 2016).

Der Zusammenhang zwischen den Kategorien der Komponente *MER-Integration* und der Aufgabenschwierigkeit wird im Regressionsmodell abgebildet ($F(1,34) = 7,701$, $p < 0,05$, $R^2 = 0,185$, $R_{kor}^2 = 0,161$, $SE=0,80$, $f=0,52$). Die Korrelation zwischen Prädiktor (MER-Integration) und der Aufgabenschwierigkeit beträgt $R = 0,43$. Demnach kann die Komponente *MER-Integration* 16,1% der Varianz der Aufgabenschwierigkeit erklären, was nach Cohen (1992) einem starken Effekt entspricht. Die Darstellung der Koeffizienten findet sich in TABELLE 35.

TABELLE 35: LINEARE REGRESSION ZUR ABHÄNGIGKEIT DER SCHWIERIGKEIT VON DER KOMPONENTE MER-INTEGRATION (ALS DUMMY-VARIABLE IN DAS MODELL EINBEZOGEN)

Komponenten	nicht standardisierte Koeffizienten		standardisierte Koeffizienten β	T	p
	B	SE			
Regressionskonstante B_0	0,049	0,179		0,272	0,787
MERII-Integration	0,745	0,268	0,430	2,775	0,009
Varianzaufklärung: $R^2 = 0,19$, $R_{kor}^2 = 0,16$ ($p < 0,05$)					

Die MERI-Integration wird als Referenzgruppe definiert und wird als Regressionskonstante B_0 ausgegeben ($B_0 = 0,049$). Die Konstante entspricht in der vorliegenden Analyse der postulierten Schwierigkeit einer maximal einfachen Aufgabe (vgl. Wellnitz, 2012, S. 118). Liegt keines der schwierigkeitsbestimmenden Aufgabenmerkmale vor, kann die Schwierigkeit dieser Aufgabe vorhergesagt werden (Hartig, 2007). Die nicht standardisierten Koeffizienten (Regressionsgewichte B) geben an, um wie viele Logits eine Testaufgabe mit diesem Merkmal (hier MERI-Integration) leichter oder schwieriger ist als eine Testaufgabe ohne dieses Merkmal (vgl. Wellnitz, 2012). Dies bedeutet, dass eine Aufgabe der *MERII-Integration* um 0,75 Logits schwieriger ist, wenn die Aufgabe eine Konstruktionsnahe Integration (MERII) erfordert.

9.3.2 AUFGABENKONTEXT UND REPRÄSENTATIONSTYP

Aufgrund der unterschiedlich verteilten Testaufgaben auf die einzelnen Aufgabenkontexte (Ökologie und Genetik mit jeweils neun Testitems, Stoffwechsel mit 18 Testitems) wurde zur Adjustierung der F-Werte der Welch-Test verwendet. Die einfaktorielle Varianzanalyse zur Ermittlung von Schwierigkeitsunterschieden in den drei untersuchten Aufgabenkontexten (Ökologie, Stoffwechsel, Genetik) führte zu keinen signifikanten Ergebnissen ($F(2,33) = 1,139$, $p = 0,332$). Dabei sind Aufgaben aus dem Bereich der Stoffwechselphysiologie für die untersuchte Stichprobe am schwierigsten zu lösen ($M_{Stw}=0,54$, $SD=0,81$), gefolgt von Aufgaben aus dem Bereich der Ökologie ($M_{Öko}=0,42$, $SD=0,97$). Am leichtesten waren Aufgaben zum Inhalt Genetik ($M_{Gen}=0,01$, $SD=0,89$). In ABBILDUNG 36 ist die Verteilung der Itemschwierigkeiten für Aufgaben aus den drei verschiedenen Kontexten dargestellt. Die mit einem Stern markierten Testaufgaben (Item See09 und Öko02) zeigen extreme Werte, d.h. der Abstand vom 25%- bzw. 75%-Perzentil ist um mehr als das Dreifache vom Perzentil entfernt.

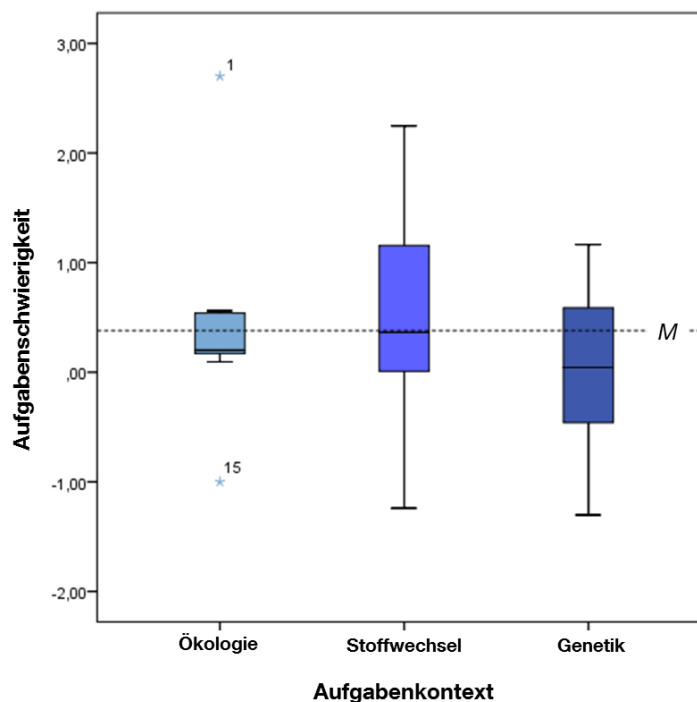


ABBILDUNG 36: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH AUFGABENKONTEXT (N=36 ITEMS, M = MITTELWERT)

Hinsichtlich des Repräsentationstyps (vgl. ABBILDUNG 37) wurde zur Berechnung der Schwierigkeitsunterschiede ein t-Test für unabhängige Stichproben durchgeführt. Dieser setzt Varianzhomogenität voraus, die überprüft wurde. Der Levene-Test ist nicht signifikant ($F(4,45) = 0,188$, $p = 0,83$). Die Teststatistik beträgt $t(34) = 1,324$ ($p = 0,194$). Damit liegt kein signifikanter Unterschied zwischen Diagramm- und Schemaaufgaben vor, wobei Diagramme schwieriger zu lösen sind ($M_D=0,57$, $SD=0,23$) als Schemazeichnungen ($M_S=0,19$, $SD=0,17$).

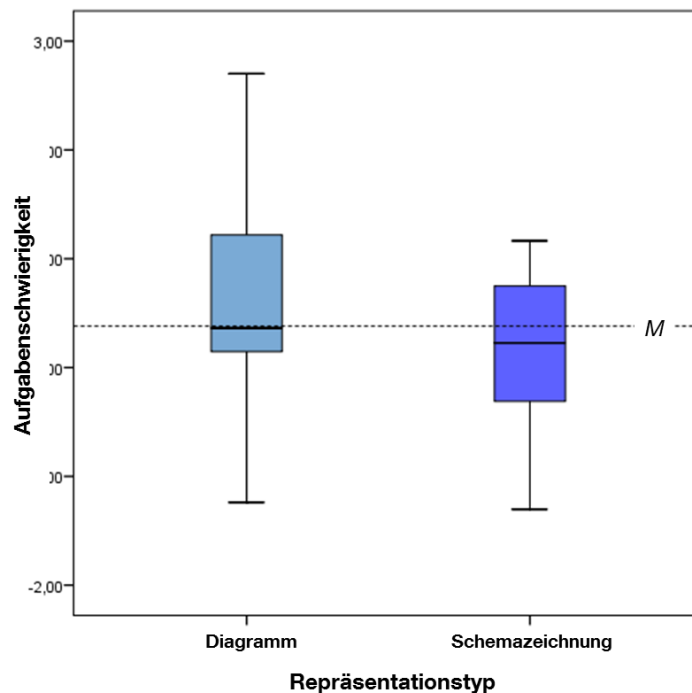


ABBILDUNG 37: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH REPRÄSENTATIONSTYP (N=36 ITEMS)

9.3.3 SCHWIERIGKEITSGENERIERENDE MERKMALE

Im Rahmen der postulierten Annahmen, dass es sich bei Repräsentationskompetenz um ein mehrdimensionales Modell handelt, das neben der MER-Integration auch die Aufgabenkontexte und den Repräsentationstyp berücksichtigt (s. Kapitel 3.3), werden neben der einfaktoriellen Varianzanalyse zusätzlich mehrfaktorielle Analysen sowie multiple lineare Regressionen durchgeführt.

Im ersten Schritt wurde sichergestellt, dass Varianzhomogenität vorliegt. Der Levene-Test ist nicht signifikant ($F(5,30) = 0,936, p = 0,472$), d.h. die Fehlervarianz der abhängigen Variable (Itemschwierigkeit) ist über die Faktoren *MER-Integration* und *Aufgabenkontext* hinweg gleich. Die Ergebnisse zeigen, dass MERII-Aufgaben ($M_{MERII}=0,79, SD=0,71$) schwieriger zu lösen sind als MERI-Aufgaben ($M_{MERI}=0,05, SD=0,87$). Am schwierigsten zu lösen sind MERII-Aufgaben aus dem Kontext Ökologie ($M_{MERII_Öko}=1,04, SD=1,11$) und der Stoffwechselphysiologie ($M_{MERII_Stw}=0,92, SD=0,49$). Aufgaben aus dem Themenbereich Genetik zeigen innerhalb der MERII-Integration einen Wert von $M_{MERII_Gen}=0,50 (SD=0,59)$. MERI-Aufgaben werden insgesamt leichter gelöst. Dies trifft vor allem auf MERI-Aufgaben der Genetik zu ($M_{MERI_Gen}=-0,97, SD=0,45$). Ökologie-Aufgaben sind ebenfalls eher leicht für die vorliegende Stichprobe zu lösen ($M_{MERI_Öko}=-0,70, SD=0,52$), während Aufgaben zur Informationsentnahmenahen Integration (MERI) in der Stoffwechselphysiologie schwieriger zu lösen sind ($M_{MERI_Stw}=0,35, SD=0,88$). Der Test zur zweifaktoriellen Varianzanalyse ist signifikant, d.h. das Gesamtmodell

ist signifikant ($F(5,30) = 3,636, p < 0,05, \eta_p^2=0,38$). Im Hinblick auf die Modellgüte liegt für das korrigierte R-Quadrat ein Wert von $R^2_{\text{kor.}} = 0,27$ vor, d.h. insgesamt werden 27,4 % Varianz aufgeklärt. Insgesamt liegt ein Haupteffekt der *MER-Integration* ($F(1,30) = 14,915, p < 0,01, \eta_p^2=0,33$) und des *Aufgabenkontexts* ($F(2,30) = 3,764, p < 0,05, \eta_p^2=0,20$) auf die Itemschwierigkeit vor. Die beiden Effekte liegen im mittleren bis starken Bereich (vgl. Cohen, 1988). Der Interaktionsterm von *MER-Integration* und *Aufgabenkontext* auf die Itemschwierigkeit ist nicht signifikant ($F(2,30) = 1,053, p = 0,362$), d.h. es liegt kein Interaktionseffekt zwischen den unabhängigen Variablen vor. Hinsichtlich des Aufgabenkontexts wurde zusätzlich überprüft, inwieweit sich die drei Faktorstufen Ökologie, Stoffwechsel und Genetik unterscheiden. Hierzu werden Berechnungen von Post-hoc-Tests zu Mehrfach-Vergleichen mit Bonferroni-Korrektur durchgeführt. In diesem Zusammenhang zeigen die Ergebnisse keine signifikanten Unterschiede zwischen den drei Aufgabenkontexten ($p > 0,05$).

Bezieht man in diese zweifaktorielle Analyse den Repräsentationstyp als dritten Faktor in das Modell ein, lassen sich zusammenfassend folgende Schwierigkeiten diagnostizieren (vgl. TABELLE 36).

TABELLE 36: VERGLEICH DER MITTLEREN ITEMSCHWIERIGKEITEN NACH KOMPONENTEN DER MER-INTEGRATION, GETRENNT NACH AUFGABENKONTEXTEN UND DEM REPRÄSENTATIONSTYP (N=36 ITEMS)

		MER I			MER II		
		N	M	SD	N	M	SD
Ökologie	Diagramm	6	-0,73	0,52	3	1,04	1,11
	Schema	–	–	–	–	–	–
Stoffwechsel	Diagramm	6	0,56	1,20	3	1,02	0,59
	Schema	6	0,14	0,39	3	0,82	0,46
Genetik	Diagramm	–	–	–	–	–	–
	Schema	3	-0,97	0,45	6	0,50	0,59

Insgesamt sind Diagrammaufgaben zur MERII-Integration ($M_{\text{MERII}_D}=1,04, SD=0,85$) schwieriger als Diagrammaufgaben zur MERI-Integration ($M_{\text{MERI}_D}=0,27, SD=0,97$). Schemaaufgaben zeigen dabei die folgenden mittleren Itemschwierigkeiten von $M_{\text{MERI}_S}=-0,23 (SD=0,67)$ und von $M_{\text{MERII}_S}=0,60 (SD=0,54)$, wobei auch hier Aufgaben zur Konstruktionsnahen Integration (MERII) schwieriger zu lösen sind als Aufgaben der Informationsentnahmenahen Integration

(MERI). Wird zusätzlich der Aufgabenkontext berücksichtigt, kann festgehalten werden, dass Aufgaben zur Diagrammkonstruktion in der Ökologie und Stoffwechselphysiologie insgesamt am schwierigsten zu lösen sind. Bei Schemaaufgaben sind Genetikaufgaben, die eine Informationsentnahmenähe Integration (MERI) verlangen, insgesamt sehr leicht zu lösen sind, während Schemaaufgaben aus dem Bereich der Stoffwechselphysiologie, die eine MERII-Integration verlangen, eher schwierig für die untersuchte Stichprobe zu lösen waren. Im Rahmen der mehrfaktoriellen Varianzanalyse wurde erneut die Voraussetzung der Varianzhomogenität überprüft und sichergestellt (Levene-Test $F(7,28) = 1,962, p = 0,097$). Der Test zur dreifaktoriellen Varianzanalyse ist signifikant, d.h. das Gesamtmodell ist signifikant ($F(7,30) = 2,662, p < 0,05, \eta_p^2=0,40$). Im Hinblick auf die Modellgüte liegt für das korrigierte R-Quadrat ein Wert von 0,25 vor, d.h. insgesamt werden 25% Varianz aufgeklärt. Für das Modell liegt ein Haupteffekt der *MER-Integration* ($F(1,28) = 13,966, p < 0,01, \eta_p^2=0,33$) auf die Itemschwierigkeit vor. Hinsichtlich des Aufgabenkontexts und des Repräsentationstyps können keine Haupteffekte auf die Itemschwierigkeiten festgestellt werden ($p > 0,05$). Auch die Interaktionsterme von *MER-Integration* und Aufgabenkontext bzw. Repräsentationstyp und von Aufgabenkontext und Repräsentationstyp zeigen keine statistisch signifikanten Effekte.

Aufgrund der Annahme, dass sich die verschiedenen Komponenten des Kompetenzmodells auf die Aufgabenschwierigkeit auswirken, wird eine multiple Regression auf Itemparameter mit Dummy-Kodierung durchgeführt. Als abhängige Variable wird die Itemschwierigkeit (intervallskaliert) definiert, die unabhängigen Variablen sind die Kategorien der Komponenten (nominalskaliert). Damit wird die Vorhersage der Aufgabenschwierigkeit durch die verschiedenen Komponenten des postulierten Kompetenzmodells überprüft. In einem ersten Schritt gehen die unabhängigen Variablen *MER-Integration* und *Aufgabenkontext* in das Regressionsmodell ein. Der Zusammenhang zwischen den Kategorien der einzelnen Dimensionen (*MER-Integration* und *Aufgabenkontext*) und Komponenten sowie der Aufgabenschwierigkeit wird im Regressionsmodell abgebildet ($F(3,32) = 5,340, p < 0,05, R^2 = 0,33, R^2_{\text{kor}} = 0,271, SE=0,75, f=0,71$). Demnach klären die Komponenten der beiden Dimensionen 27,1 % Varianz in der Aufgabenschwierigkeit auf, was nach Cohen (1992) einem starken Effekt entspricht. In TABELLE 37 sind die Ergebnisse der multiplen linearen Regression dargestellt. Die Komponenten *MERI-Integration* und *Stoffwechsel* bleiben wegen ihrer statistischen Abhängigkeit von den anderen Komponenten unberücksichtigt. Ihre Schwierigkeit ergibt sich aus der Regressionsgleichung.

TABELLE 37: MULTIPLE LINEARE REGRESSION ZUR ABHÄNGIGKEIT DER SCHWIERIGKEIT VON DEN DIMENSIONEN MER-INTEGRATION MIT DEN KOMPONENTEN MERI UND MERII SOWIE VON DER DIMENSION AUFGABENKONTEXT MIT DEN KOMPONENTEN ÖKOLOGIE, STOFFWECHSEL UND GENETIK (ALS DUMMY-VARIABLE IN DAS MODELL EINBEZOGEN)

Komponenten	nicht standardisierte Koeffizienten		standardisierte Koeffizienten β	T	p
	B	SE			
Regressionskonstante B_0	0,231	0,196		1,179	0,247
MERII-Integration	0,935	0,260	0,539	3,594	0,001
Ökologie	-0,223	0,306	-0,112	-0,728	0,472
Genetik	-0,844	0,316	-0,424	-2,666	0,012

Varianzaufklärung: $R^2 = 0,33$, $R^2_{\text{kor}} = 0,27$ ($p < 0,05$)

Mit Ausnahme der Aufgabenmerkmale Ökologie und Genetik sind die Standardfehler der nicht standardisierten Koeffizienten bei allen Prädiktoren wünschenswert kleiner als die Absolutwerte der nicht standardisierten Regressionsgewichte B . Die Regressionsgewichte zeigen, dass Testaufgaben zur MERII-Integration schwieriger zu lösen sind als Testaufgaben zur MERI-Integration beziehungsweise, dass die Aufgabenkontexte Ökologie und Genetik leichter gelöst werden als Testaufgaben aus der Stoffwechselphysiologie. Der relative Einfluss der verschiedenen Aufgabenmerkmale auf die Itemschwierigkeit kann über die standardisierten Koeffizienten (Regressionsgewichte β) verdeutlicht werden (vgl. Wellnitz, 2012, S. 119). Je höher die Werte von β sind, desto mehr trägt dieses Aufgabenmerkmal zur Aufklärung der Streuung der Itemschwierigkeit bei. So liefert das Merkmal Konstruktionsnahe Integration (MERII) mit $\beta = 0,54$ ($p < 0,01$) einen hoch signifikanten Beitrag zur Varianzaufklärung. Für die Dimension der Aufgabenkontexte kann festgestellt werden, dass nur der Kontext Genetik einen signifikanten Einfluss auf die Itemschwierigkeit vorhersagt. Dies gilt nicht für den Kontext Ökologie. Insgesamt werden mit den zwei Dimensionen MER-Integration und Aufgabenkontext 33% der Gesamtvarianz aufgeklärt. Unter Berücksichtigung des Repräsentationstyps als weitere unabhängige Variable und damit Prädiktor zur Vorhersage der Aufgabenschwierigkeit konnte keine Steigerung der Varianzaufklärung mehr erreicht werden ($F(4,31) = 4,250$, $p < 0,01$, $R^2 = 0,35$, $R^2_{\text{kor}} = 0,271$). Des Weiteren zeigen sich keine statistisch bedeutsamen Effekte hinsichtlich der Einflussstärke auf die Schwierigkeit, weshalb die Ergebnisse in diesem Fall stark reduziert festgehalten werden. Die Erwartung, dass neben der *MER-Integration* auch die Konstrukte *Aufgabenkontexte* und der *Repräsentationstyp* einen

Einfluss haben, kann damit nur bedingt bestätigt werden. Gleichzeitig sollten aufgrund der geringen Itembesetzung innerhalb der einzelnen Zellen die Ergebnisse vorsichtig interpretiert werden.

9.4 PRÜFUNG VON KOMPETENZUNTERSCHIEDEN

Im Folgenden werden die Fähigkeiten der Studierenden aus unterschiedlichen Studiengängen in den Teilkompetenzen *MERI* und *MERII* sowie in den Teilkomponenten des *Aufgabenkontexts* (Ökologie, Stoffwechsel, Genetik) und des *Repräsentationstyps* (Diagramm, Schema) verglichen. Zu diesem Zweck werden die auf Basis aller 36 Items geschätzten Personenfähigkeiten aller Studierenden (N=548) herangezogen. Im Anschluss wird untersucht, welche Zusammenhänge zwischen den Personenmerkmalen und der Repräsentationskompetenz bestehen.

Die zugrundeliegenden Werte für die Personenfähigkeiten auf Basis aller Items können aus der Berechnung für die itemzentrierte, zweidimensionale Skalierung zur *MER-Integration* (*MERI* und *MERII*) entnommen werden. Dabei werden pro Testperson zwei Plausible Values (PVs) aus dem 2D-MER-Modell verwendet. Für die weiteren Berechnungen zur Prüfung von Kompetenzunterschieden wurde analog vorgegangen. Dabei wurden für die Dimension *Aufgabenkontexte* drei PVs aus dem CON-3D-Modell herangezogen und für die Dimension *Repräsentationstyp* zwei PVs aus dem TYP-2D-Modell. Um zu prüfen, ob die jeweiligen Komponenten der Dimensionen *MER-Integration*, der *Aufgabenkontext* und der *Repräsentationstyp* in der Grundgesamtheit einen signifikanten Einfluss haben, wird eine einfaktorielle Varianzanalyse mit Messwiederholung durchgeführt.

9.4.1 KOMPETENZAUSPRÄGUNGEN DER TEILKOMPETENZEN IM VERGLEICH

Die Dimension *MER-Integration* wird für die Gesamtstichprobe (N=548) auf mögliche Leistungsunterschiede geprüft. Im Rahmen der Teilkompetenzen *MERI* und *MERII* kann festgehalten werden, dass zwischen den Teilkompetenzen *MERI* und *MERII* signifikante Unterschiede in den Personenfähigkeiten bestehen ($z = -8,783$, $p < 0,001$, $n=545$). Zur Berechnung der Kompetenzunterschiede wurde der nicht-parametrische Wilcoxon-Test durchgeführt, da die Voraussetzung der Daten auf Normalverteilung nicht erfüllt ist (Shapiro-Wilk-Test $p < 0,05$). Die mittlere Personenfähigkeit liegt bei $M_{\theta} = -0,34$ ($SD=1,05$). Die mittleren Personenfähigkeiten für die *Informationsentnahmenähe Integration* (*MERI*) beträgt $M_{MERI} = -0,03$ ($SD=0,05$) und für die *Konstruktionsnähe Integration* (*MERII*) $M_{MERII} = -0,65$ ($SD=0,05$). In ABBILDUNG 38 werden die Leistungsunterschiede graphisch veranschaulicht.

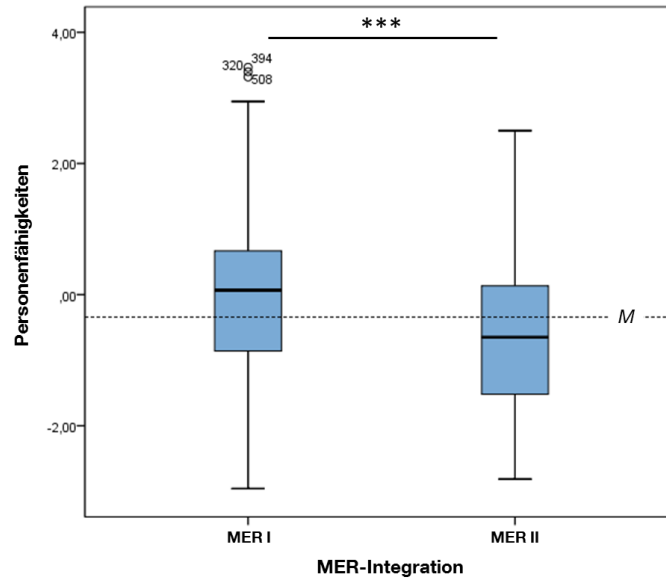


ABBILDUNG 38: PERSONENFÄHIGKEITEN BEI DER MER-INTEGRATION MIT MERI: INFORMATIONSENTNAHMENAEHE INTEGRATION UND MERII: KONSTRUKTIONSNAAHE INTEGRATION (M = MITTELWERT)

Zur Überprüfung, ob zwischen den beiden *Repräsentationstypen Diagramm* und *Schemazeichnung* ebenfalls signifikante Unterschiede in den Personenfähigkeiten vorliegen, wurde erneut der Wilcoxon-Test durchgeführt, da die Voraussetzung auf Normalverteilung nach dem Shapiro-Wilk-Test nicht gegeben sind ($p < 0,05$). Die mittlere Personenfähigkeit für Diagrammaufgaben liegt bei $M_D = -0,52$ ($SD = 0,05$) und für Schemaaufgaben bei $M_S = -0,19$ ($SD = 0,05$). Demnach zeigt sich über die Repräsentationstypen ein höchst signifikanter Einfluss auf die Personenfähigkeiten ($z = -4.574$, $p < 0,001$, $n = 547$). Die graphisch veranschaulichten Leistungsunterschiede können ABBILDUNG 39 entnommen werden.

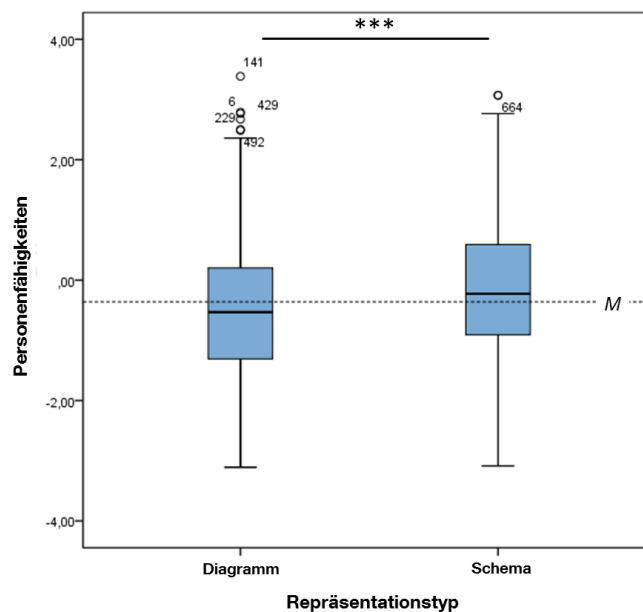


ABBILDUNG 39: PERSONENFÄHIGKEITEN BEIM REPRÄSENTATIONSTYP DIAGRAMM UND SCHEMA (M = MITTELWERT)

Um zu prüfen, ob die Aufgabenkontexte Ökologie, Stoffwechsel und Genetik in der Grundgesamtheit einen signifikanten Einfluss auf die Personenfähigkeiten der Studierenden haben, wird eine einfaktorielle Varianzanalyse mit Messwiederholung durchgeführt. Der Mauchly-Test auf Sphärizität ist nicht signifikant ($Mauchly-W(2) = 0,998, p = 0,647$) und bestätigt die Voraussetzung der Varianzhomogenität zwischen den Personenfähigkeiten. Da für die einzelnen Aufgabenkontexte keine Normalverteilung vorliegt ($p < 0,05$), werden die Analysen mithilfe nicht-parametrischer Verfahren durchgeführt. Dafür wird der Unterschied in den Personenfähigkeiten mit dem Friedman-Test geprüft. Dieser zeigt, dass signifikante Unterschiede in den Leistungen der Studierenden hinsichtlich des Aufgabenkontextes vorliegen ($\chi^2(2) = 12,695, p < 0,001, n=521$). Zusätzlich durchgeführt multivariate Tests bestätigen einen schwachen Haupteffekt des Aufgabenkontextes ($V = 0,05, F(2,1040) = 13,912, p < 0,001, \eta_{part.}^2=0,05$). Die mittleren Personenfähigkeiten betragen für den *Aufgabenkontext* *Ökologie* $M_{Öko}=-0,36 (SD=0,05)$, für den Kontext *Stoffwechsel* $M_{Stw}=-0,44 (SD=0,06)$ sowie für den Kontext *Genetik* $M_{Gen}=-0,09 (SD=0,05)$. Um zu klären, welche Personenkenwerte sich im Mittel signifikant voneinander unterscheiden, werden anschließend paarweise Vergleiche nach Bonferroni durchgeführt (Field, 2009b). In ABBILDUNG 40 sind die Leistungsunterschiede der drei verschiedenen *Aufgabenkontexte* graphisch veranschaulicht.

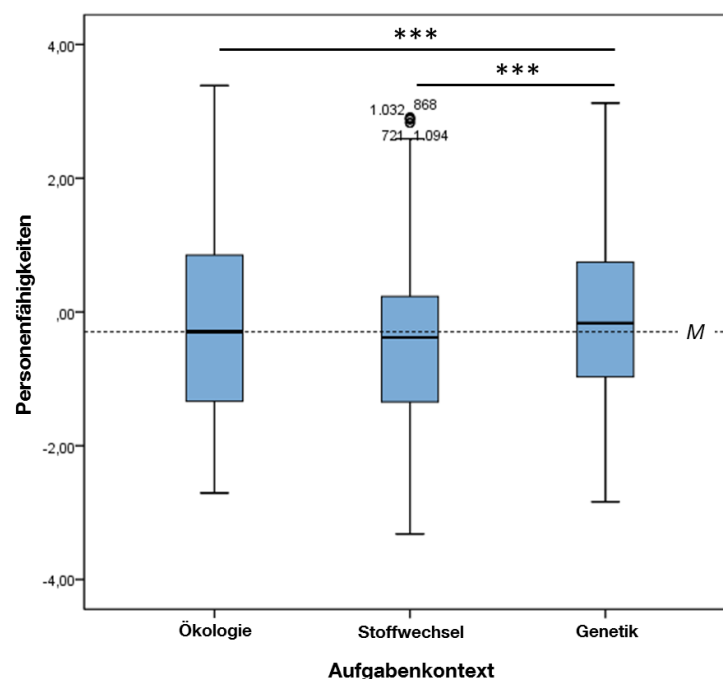


ABBILDUNG 40: PERSONENFÄHIGKEITEN BEI DEN AUFGABENKONTEXTEN ÖKOLOGIE, STOFFWECHSEL, GENETIK (M = MITTELWERT)

Die Prüfung mit dem Post-Hoc-Test zeigt, dass sich die mittleren Leistungsdifferenzen zwischen den Aufgabenkontexten *Ökologie* und *Genetik* ($-0,27, 95\%-CI [-0,8, -0,10], p < 0,001$) sowie zwischen *Stoffwechsel* und *Genetik* ($-0,36, 95\%-CI [-0,53, -0,18], p < 0,001$) signifikant

voneinander unterscheiden. Die Unterschiede in den Personenfähigkeiten zwischen den Kontexten *Ökologie* und *Stoffwechsel* sind erwartungswidrig hingegen statistisch nicht bedeutsam.

9.4.2 STUDIENGANG

Die Kompetenzaspekte *Informationsentnahmenähe Integration (MERI)* und *Konstruktionsnähe Integration (MERII)* werden vor dem Hintergrund des Studiengangs auf mögliche Leistungsunterschiede überprüft. Dies wird für die *MER-Integration* nach *MERI* und *MERII* vorgenommen. Bei der Analyse wird zwischen fünf Studiengängen differenziert: Lehramt Biologie, Lehramt Mathematik (und Biologie), Biologie, Ingenieurwissenschaften, Mathematik. Studierende, die keine Angabe bezüglich des Studiengangs gemacht haben, wurden von der Analyse ausgeschlossen. Die Analysen können aufgrund der Annahme auf Normalverteilung der Daten ($p > 0,05$) mit parametrischen Testverfahren durchgeführt werden. Dabei weisen die mittleren Fähigkeitsparameter auf Kompetenzunterschiede zwischen den Studierenden hin (vgl. TABELLE 38).

TABELLE 38: VERGLEICH DER STUDIERENDENFÄHIGKEITEN IN DEN KOMPETENZEN MERI UND MERII DIFFERENZIERT NACH STUDIENGANG (N=533)

	MERI			MER II	
	N	M	SD	M	SD
Lehramt Biologie	84	-0,15	1,27	-0,55	1,30
Lehramt Mathematik (und Biologie)	35	-0,18	1,20	-0,79	1,16
Biologie	145	0,06	1,22	-0,47	1,06
Ingenieurwissenschaften	168	-0,16	1,15	-0,89	1,22
Mathematik	101	0,14	1,23	-0,59	1,25

Die mittlere Repräsentationskompetenz nimmt im Rahmen der *MERI-Integration* in der Reihenfolge $M_{\text{Mathematik}} > M_{\text{Biologie}} > M_{\text{Lehramt Biologie}} > M_{\text{Ingenieurwissenschaft}} > M_{\text{Lehramt Mathematik (und Biologie)}}$ und im Rahmen der *MERII-Integration* in der Reihenfolge $M_{\text{Biologie}} > M_{\text{Lehramt Biologie}} > M_{\text{Mathematik}} > M_{\text{Lehramt Mathematik (und Biologie)}} > M_{\text{Ingenieurwissenschaft}}$ ab. Die Unterschiede unter Berücksichtigung von MERI und MERII sind ABBILDUNG 41 und ABBILDUNG 42 graphisch veranschaulicht.

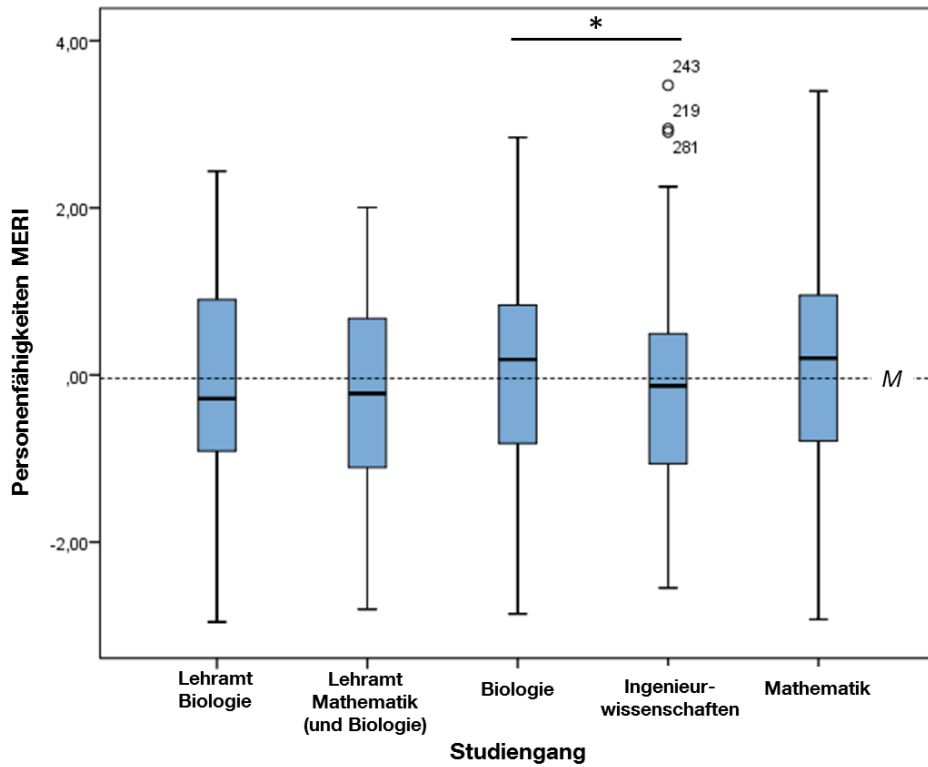


ABBILDUNG 41: PERSONENFÄHIGKEITEN DER MERI-INTEGRATION NACH STUDIENGANG (N=533, $M =$ MITTELWERT)

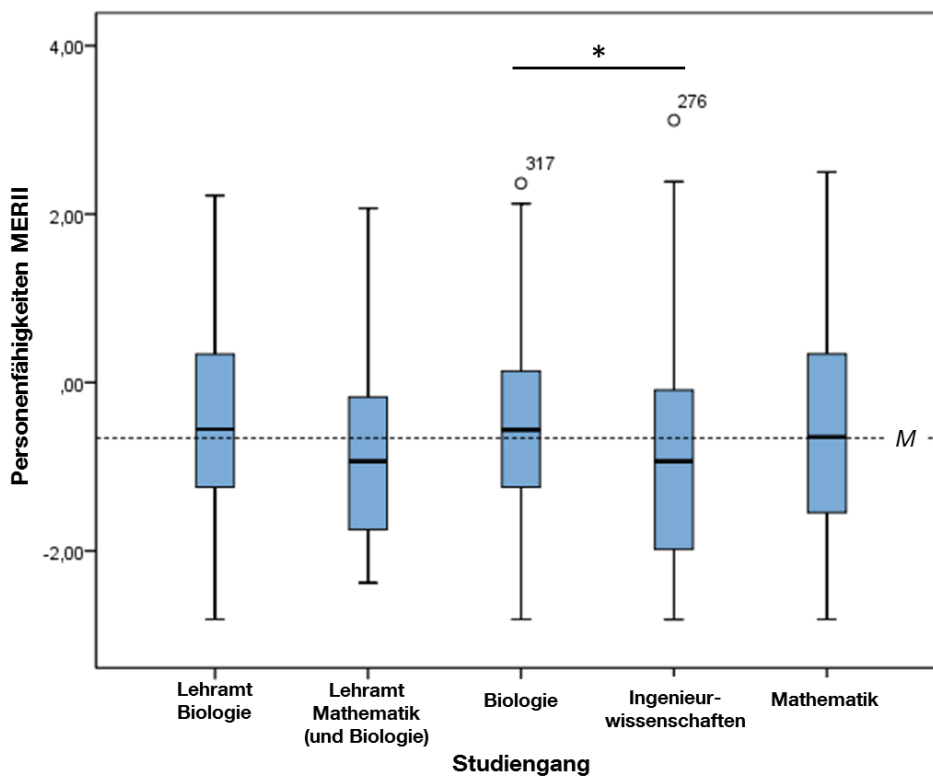


ABBILDUNG 42: PERSONENFÄHIGKEITEN DER MERII-INTEGRATION NACH STUDIENGANG (N=533, $M =$ MITTELWERT)

Aufgrund der Sphärizitätsverletzung im Mauchly-Test wird die Korrektur nach Greenhouse-Geisser für die Überprüfung herangezogen. Die Mittelwertsunterschiede der Personenfähigkeiten sind in der Grundgesamtheit *Repräsentationskompetenz* zwischen den Studiengängen signifikant ($F(1,4) = 58,783$, $p < 0,001$, $\eta_{part.}^2=0,10$), was nach Cohen (1988) einem starken Effekt entspricht. Es gibt also einen Haupteffekt des Studiengangs auf die *MER-Integration*. Zur Ermittlung, welche Leistungen der Studierenden sich spezifisch nach dem Studiengang voneinander unterscheiden, werden paarweise Vergleiche durchgeführt. Aufgrund der stark divergierenden Stichprobenumfänge in den Studiengängen wird der Post-Hoc-Test nach Hochberg (GT2) herangezogen. Die paarweisen Vergleiche im Rahmen der *MER-Integration* zeigen, dass sich die Personenfähigkeiten der Studierenden im Studiengang Biologie ($M_{Biologie}=-0,37$, $SD=0,13$) signifikant von denen der Studierenden im Studiengang Ingenieurwissenschaften ($M_{Ingenieurwissenschaften}=-0,55$, $SD=0,07$) unterscheiden ($0,37$, 95%-CI $[0,04, 0,70]$, $p < 0,05$). Im Hinblick auf die drei untersuchten *Aufgabenkontexte* zeigen sich ebenfalls signifikante Unterschiede zwischen diesen beiden Studierendengruppen im Kontext *Stoffwechsel* ($0,42$, 95%-CI $[0,04, 0,83]$, $p < 0,05$), während signifikante Unterschiede im Kontext *Genetik* zwischen Studierenden im Studiengang Ingenieurwissenschaften und Mathematik diagnostiziert werden konnten ($-0,46$, 95%-CI $[-0,89, -0,03]$, $p < 0,05$). Hervorzuheben ist, dass Mathematikstudierende den höchsten Fähigkeitswert bei den Testaufgaben in der *Genetik* aufweisen ($M_{Genetik_Mathematiker}=0,22$, $SD=1,16$), in den beiden anderen Kontexten *Stoffwechsel* und *Ökologie*, zeigen Studierende im Studiengang Biologie die höchsten Fähigkeitswerte ($M_{Stoffwechsel_Biologen}=-0,21$, $SD=1,20$; $M_{Ökologie_Biologen}=-0,23$, $SD=1,22$). Alle weiteren Mittelwertsunterschiede ergeben keine statistisch bedeutsamen Leistungsunterschiede zwischen den einzelnen Studiengängen.

Zur Stützung dieser Ergebnisse wurden auf latenter Ebene zusätzlich die unstandardisierten Regressionsgewichte berechnet. Die Normalverteilung der Residuen als Voraussetzung (vgl. Field, 2009b) wird durch eine grafische Testung bestätigt. Damit können die Regressionsmodelle als Ganzes und die einzelnen Regressionskoeffizienten auf ihre statistische Signifikanz überprüft werden. Die Komponente *Studiengang Ingenieurwissenschaften* wurde als Referenzgruppe definiert und wird dementsprechend als Regressionskonstante ausgegeben. Zur Vorhersage der Personenfähigkeit durch die unabhängige Variable Studiengang wird ein F-Test zur Signifikanz des Regressionsmodells durchgeführt. Demnach ist das Modell als Ganzes signifikant ($F(4,531) = 3,044$, $p < 0,05$, $R^2 = 0,02$, $R^2_{korr} = 0,02$, $f=0,14$). Insgesamt können damit zwei Prozent der Streuung erklärt werden, was nach Cohen (1992) einem schwachen Effekt der Varianzaufklärung entspricht. Die Kennwerte zur multiplen Regressionsanalyse sind TABELLE 39 zu entnehmen.

TABELLE 39: MULTIPLE LINEARE REGRESSION ZUR ABHÄNGIGKEIT DER PERSONENFÄHIGKEIT VON DER KOMPONENTE STUDIENGANG (ALS DUMMY-VARIABLEN IN DAS MODELL EINBEZOGEN)

Studiengang	nicht standardisierte Koeffizienten		standardisierte Koeffizienten β	T	p
	B	SE			
Regressionskonstante B_0	-0,550	0,080		-6,880	< 0,001
Lehramt Biologie	0,184	0,139	0,064	1,324	0,186
Lehramt Mathematik (und Biologie)	0,014	0,191	0,003	0,075	0,940
Biologie	0,374	0,118	0,159	3,175	< 0,01
Mathematik	0,292	0,130	0,110	2,240	< 0,05
Varianzaufklärung: $R^2 = 0,02$, $R^2_{\text{korrt}} = 0,02$ (p < 0,05)					

Aufgrund des negativen Vorzeichens der Regressionskonstante sind die Werte dahingehend zu interpretieren, dass je positiver der nicht standardisierte Koeffizient B , desto höher sind auch die Personenfähigkeiten des jeweiligen Studiengangs. Der Kompetenzabstand zwischen Studierenden der Ingenieurwissenschaften und denen der Biowissenschaften beträgt 0,37. Niedriger werden die Kompetenzabstände im Studiengang Mathematik ($B=0.29$), Lehramt Biologie ($B=0.18$) und Lehramt Mathematik (und Biologie) ($B=0.01$). Die standardisierten Koeffizienten (β) sowie die t-Tests zeigen, dass neben dem Studiengang Biologie auch der Studiengang Mathematik eine Rolle in den Kompetenzunterschieden hat.

9.4.3 NOTEN, ABITURFACH UND GESCHLECHT

Im Rahmen der explorativen Analyse wird untersucht, welcher Zusammenhang zwischen den moderierenden Variablen Biologie- und Mathematiknote, Abiturfach sowie Geschlecht und der Personenfähigkeit besteht. Damit wird ein weiterer Beitrag zur diskriminanten Validierung des zweidimensionalen Kompetenzmodells auf die Kompetenzen *Informationsentnahme-nahe Integration (MERI)* und *Konstruktionsnahe Integration (MERII)* geleistet. Die Analysen finden über Korrelationsvergleiche statt. Damit kann geprüft werden, inwiefern die Repräsentationsfähigkeiten eine größere Nähe zu biologischen oder mathematischen Kompetenzen aufweisen. Für die untersuchte Stichprobe (N=545) liegen die Noten für die Fächer Biologie ($M=11,50$, $SD=2,16$) und Mathematik ($M=11,45$, $SD=3,01$) im gängigen Punktesystem für

Schulnoten der Oberstufe vor (s. Kapitel 4.2). Demnach handelt es sich um eine sehr homogene Gruppe in der Notenverteilung zwischen Biologie und Mathematik. Insgesamt haben 43 % der Stichprobe im Fach Biologie auch ein Abitur abgelegt, 81 % Prozent im Fach Mathematik. Zur Überprüfung der Zusammenhänge zwischen den Leistungen der Studierenden und den Schulnoten wurden Rangkorrelationen nach Spearman durchgeführt.

Die bivariate Korrelation der Personenparameter für die MERI-Integration mit der Biologienote liegt mit $r = 0,11$ ($p < 0,05$) im sehr geringen Bereich, ist aber signifikant. Ähnliche Befunde ergeben sich für die MERII-Integration ($r = 0,10$, $p < 0,05$). Im Hinblick auf das Fach Mathematik ergibt sich für die MERI-Integration eine Korrelation mit der Mathematiknote von $r = 0,03$ und für die MERII-Integration eine Korrelation von $r = 0,04$. Die Korrelationen sind statistisch nicht bedeutsam ($p > 0,05$).

Im Hinblick auf die Fragestellung, ob das Abiturfach einen Einfluss auf die Personenfähigkeit hat, liefern die Ergebnisse einen ebenfalls schwachen, aber signifikanten Effekt für die vorliegende Stichprobe ($N=548$). Die bivariate Korrelation von Personenparametern mit dem Abiturfach Biologie liegt in einem ähnlichen Wertebereich ($r = 0,12$, $p < 0,01$), ob in Mathematik auch Abitur gemacht wurde, scheint dabei keine Rolle zu spielen ($p > 0,05$).

Abschließend wurde der Vergleich der Geschlechter untersucht. Für den Gesamttest zeigt sich ein hoch signifikanter Leistungsunterschied in den Personenfähigkeiten der weiblichen Studierenden ($M=-0,24$, $SD=1,05$, $N=264$) und der männlichen Studierenden ($M=-0,47$, $SD=1,03$, $N=284$) mit $t(546) = 2,637$ ($p < 0,01$).

9.5 QUALITATIVE FEHLERANALYSE

Vor dem Hintergrund der insgesamt moderaten Reliabilitäten der einzelnen Skalen (s. Kapitel 9.1) für die ein- und mehrdimensionalen Skalierungen wird im Folgenden zusätzlich eine qualitative Fehleranalyse durchgeführt. In diesem Zusammenhang soll eine Auswahl der insgesamt 36 zur Verfügung stehenden Items reflektiert und hinsichtlich des offenen Antwortformats analysiert werden. Dabei werden einzelne markante Fehler, die bei der Beantwortung der offenen Items aufgetreten sind, vorgestellt und jeweils den im Kompetenzstrukturmodell formulierten Komponenten zugeordnet. Die vollständigen Testaufgaben, bestehend aus Aufgaben- und Itemstamm, Textinformation, Repräsentationstyp und Fragestellung können dem Anhang (s. Kapitel 14.1) entnommen werden. Im Rahmen der vorliegenden qualitativen Fehleranalyse werden ausschließlich die Schwierigkeiten durch Darstellung der Antwortmuster vorgestellt.

9.5.1 INFORMATIONSENTNAHMENAHE INTEGRATION (MERI)

In der Testaufgabe Flo08 (Aufgabenkontext Ökologie, Repräsentationstyp Diagramm) wird ein Vergleich von vier Kurven in einem Liniendiagramm erwartet. Dabei sollen die Ergebnisse diskutiert und die Frage beantwortet werden, wie sich der Einsatz des Fisches auf die Individuenzahl auswirkt. Für die Aufgabe liegen 129 Antworten aus der Hauptstudie vor. Dabei wurde das mittelschwere Item zu 54,3% richtig gelöst ($M=0,098$, $SD=0,205$). In TABELLE 40 sind typische fehlerhafte bzw. nicht vollständig gelöste Antworten beispielhaft dargestellt.

TABELLE 40: ANTWORTBEISPIELE ZUM ITEM FLO08 ZU DEN TEILKOMponentEN MERI, ÖKOLOGIE, DIAGRAMM (VGL. MUSTERLÖSUNG IM ANHANG 14.1.1)

Beispielantworten Item Flo08

„(...) Die Individuenzahl von Daphnia hat nur mit der Anwesenheit des Fisches zu tun, nicht von den anderen Arten von Plankton. Sobald mich den Fischeinsatz zurücknimmt nimmt die Individuenzahl von Daphnia allmählich ab“ (Ingenieurwissenschaften)

„Ohne Fisch ist das Ökosystem sehr stabil. (...) Wird der Fisch entnommen beginnt sich das Ökosystem wieder wie zu Beginn zu stabilisieren“ (Lehramt Mathematik, Physik)

„(...) Es zeigt das die Ceriodaphnia bevorzugt vom Fisch gefressen wird wohingegen die anderen ein gutes Zusammenleben haben“ (Lehramt Biologie, Deutsch)

Auffällig im Antwortmuster ist die Interpretation der beiden Operatoren *vergleichen* und *diskutieren*. In den meisten Fällen wurde die Aufgabe ausschließlich im Rahmen der Frage, wie sich der Fischeinsatz auf die Individuenzahl auswirkt, beantwortet. Dabei bleibt die Beantwortung auf vergleichender Ebene während dieses Fischeinsatzes. Eine schlüssige Diskussion vor dem Hintergrund der Räuber-Beute-Beziehung (Textinformation) fehlt hingegen in den meisten Fällen.

Das Item Dna02 (Aufgabenkontext Genetik, Repräsentationstyp Schema) mit einem Mittelwert von $M=-1,252$ ($SD=0,218$) war in der vorliegenden Stichprobe der Hauptstudie insgesamt am leichtesten zu lösen. Das Item wurde bei $n=123$ vorliegenden Antworten zu 75,6% richtig gelöst. Mithilfe der Code-Sonne sollte eine gegebene Kette von Aminosäuren in einen auf der RNA möglichen gespeicherten Code umgewandelt werden. Markante Fehler in der Beantwortung lagen vor, wenn die gegebene Textinformation (drei Basen stehen für eine Aminosäure, z.B. AUG für Methionin (Met)) nicht in Verbindung zur Code-Sonne gebracht werden konnten. In TABELLE 41 sind beispielhafte Antworten dargestellt.

TABELLE 41: ANTWORTBEISPIELE ZUM ITEM DNA02 ZU DEN TEILKOMPONENTEN MERI, GENETIK, SCHEMA (VGL. MUSTERLÖSUNG IM ANHANG 14.1.1)

Beispielantworten Item Dna02

(Met – Gly – Ala) – (Asn – Val – Val) – (Cys – Pro – Thr)



(Mathematik)

GUA – UGG – GCG – CAA – AUG – UGU – UCC – GCA

(Biologie)

Im zweiten Beispiel der Tabelle fehlt die Nennung des Start-Codons. Ebenfalls wurde der Code innerhalb der Code-Sonne von außen nach innen abgelesen und damit die Information aus dem Text nicht korrekt auf die Code-Sonne übertragen.

9.5.2 KONSTRUKTIONSNAHE INTEGRATION (MERII)

Die Testaufgabe See07 erfordert eine Integration aus gegebener Textinformation (Tag 1 der Probennahme mit 60 mg/L für den Zufluss), einer Wertetabelle mit Angaben zum Nitratgehalt in mg/L für Zufluss und See sowie dem Diagrammrahmen bestehend aus der gegebenen abhängigen (Tag der Probennahme) und unabhängigen Variable (Nitratgehalt in mg/L). Für diese Aufgaben liegen insgesamt n=119 Antworten vor. Das Item wurde dabei zu 37,8% richtig gelöst ($M=0.552$, $SD=0.199$). In ABBILDUNG 43 sind typische Schwierigkeiten im Rahmen der Beantwortung dargestellt.

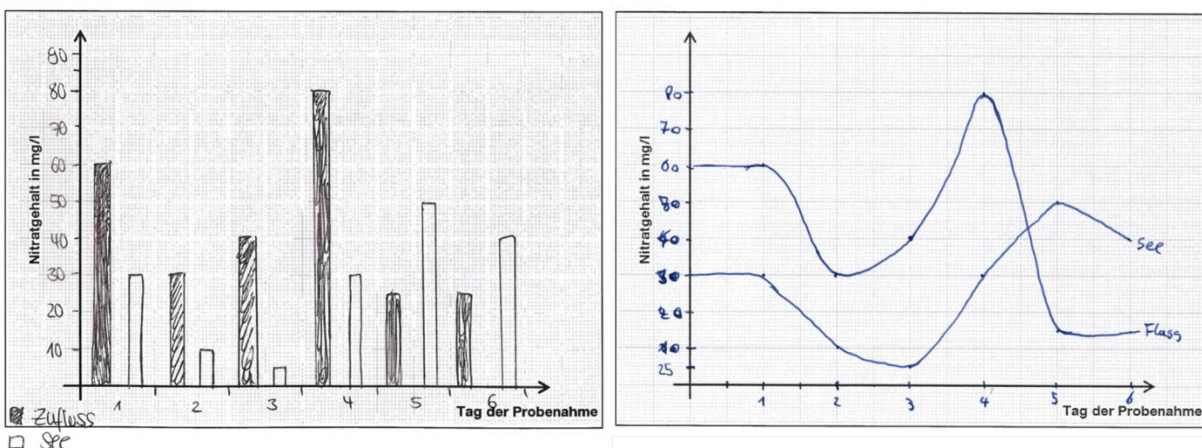


ABBILDUNG 43: ANTWORTBEISPIELE ZUM ITEM SEE07 ZU DEN TEILKOMPONENTEN MERII, ÖKOLOGIE, DIAGRAMM (VGL. MUSTERLÖSUNG IM ANHANG 14.1.1)

Wie die beiden Diagrammbeispiele in **ABBILDUNG 43** zeigen, scheint die Wahl des korrekten Diagrammtyps zu den gegebenen Werten und Variablen Schwierigkeiten bereitet zu haben. So wurde die „Zeit“ als metrisch skalierte Variable kategorial interpretiert und auf der Abszisse eingetragen. Eine weitere Schwierigkeit war die Informationsentnahme aus dem Text, dass an Tag eins der Nitratgehalt für den Zufluss 60 mg/L betrug. Diese Information wurde häufig nicht in das zu konstruierende Diagramm übertragen, sondern lediglich die Werte aus der Tabelle übernommen. Damit fehlt die Information im Diagramm, dass es sich bei dem ersten Wertepaar um den ersten Tag der Probenahme handelt.

Ein weiteres Beispiel aus der Teilkompetenz MERII ist die Testaufgabe Epo05 (Aufgabenkontext Stoffwechsel, Repräsentationstyp Diagramm). Dabei handelt es sich um ein tendenziell schwierigeres Item mit einem Mittelwert von $M=1,282$ ($SD=0,237$) und einer Lösungshäufigkeit von lediglich 23,1% bei insgesamt $n=108$ vorliegenden Antworten. Zur Lösung der Aufgabe ist es erforderlich den gegebenen Text sowie das dargestellte Diagramm zu integrieren, damit im Anschluss daran ein neues Diagramm gezeichnet werden kann, das den Hämokrit-Wert in Abhängigkeit von der EPO-Dosierung nach einer Behandlungsdauer von sechs Wochen zeigt. Die Schwierigkeit der Aufgabe scheint vor allem im Bereich der Werteentnahme aus dem gegebenen Diagramm zu bestehen. Dabei treten einerseits Fehler im AbleSEN der Werte im gegebenen Diagramm zum gefragten Zeitpunkt (nach einer Behandlungsdauer sechs Wochen) sowie der anschließenden richtigen Eintragung dieser in das neue Diagramm (vgl. **ABBILDUNG 44**).

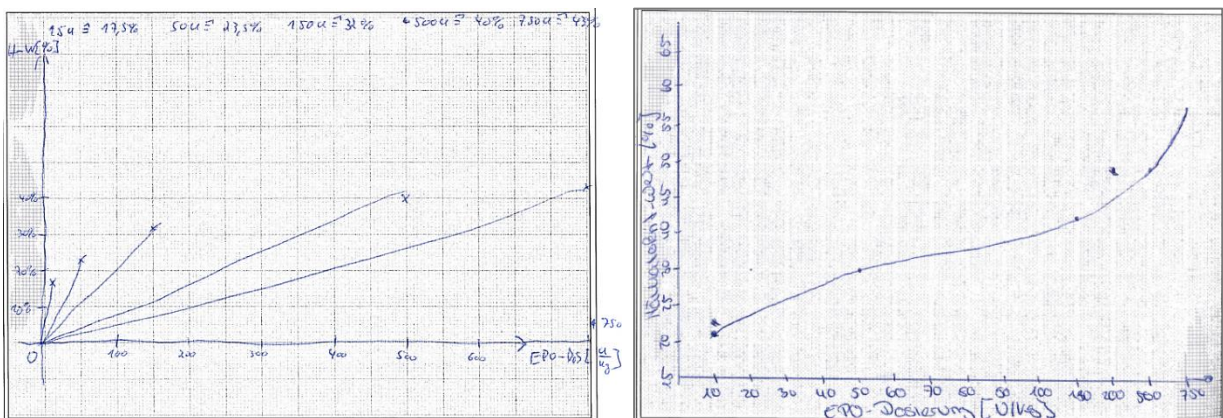


ABBILDUNG 44: ANTWORTBEISPIELE ZUM ITEM EPO05 - WERTE ENTNEHMEN UND EINTRAGEN (MERII, ÖKOLOGIE, DIAGRAMM)

Andererseits scheint es Schwierigkeiten in der Wahl des korrekten Diagrammtyps sowie bei der korrekten Zuordnung der abhängigen und unabhängigen Variablen auf der Ordinate und Abszisse zu geben (vgl. **ABBILDUNG 45**, **ABBILDUNG 46**).

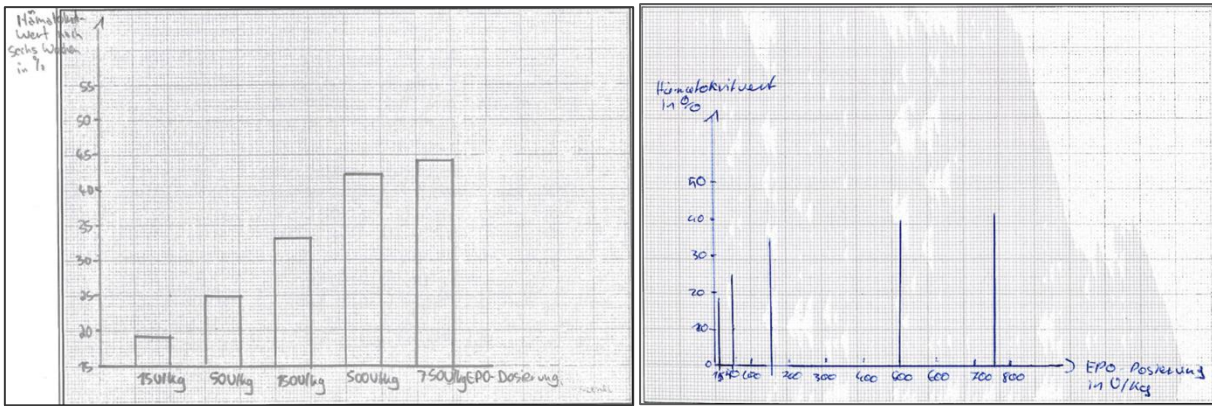


ABBILDUNG 45: ANTWORTBEISPIELE ZUM ITEM EPO05 – DIAGRAMMTYP

Wie das Beispiel in ABBILDUNG 46 zeigt, bestehen ebenfalls Schwierigkeiten in der Wahl einer angemessenen Skalierung, die im Rahmen des gegebenen Diagramms hätten exakter ausgewählt werden können.

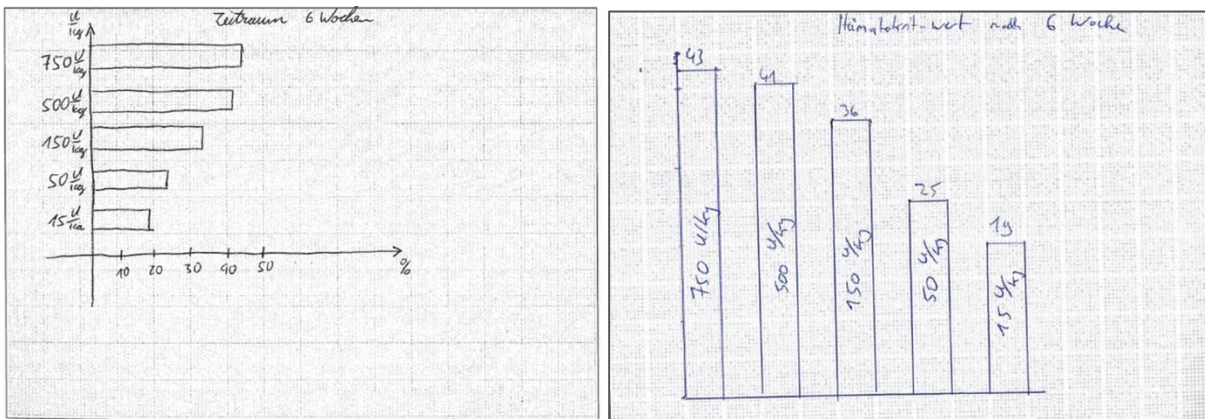


ABBILDUNG 46: ANTWORTBEISPIELE ZUM ITEM EPO05 – ACHSENZUORDNUNG UND SKALIERUNG

Im Folgenden werden die Schwierigkeiten beim Lösen von Testaufgabe Blu05 (Aufgabenkontext Genetik, Repräsentationstyp Schema) vorgestellt. Mit einer Lösungshäufigkeit von 41.8% bei $n=122$ vorliegenden Antworten und einem Mittelwert von $M=0,363$ ($SD=0,193$) handelt es sich um ein Item im Bereich der mittleren Schwierigkeit dieser Untersuchung. Zur korrekten Lösung der Aufgaben soll die gegebene Tabelle mit vervollständigt werden. In ABBILDUNG 47 sind die typischen Fehler und Schwierigkeiten bei der Beantwortung der Testaufgabe dargestellt.

Antigene (Erythrozyten)	A	B		0	
Antikörper (Serum, der wässrige Bestandteil des Blutes)			keine		A / B
Blutgruppe	A	B	A/B	0	

Antigene (Erythrozyten)	A	B		0	
Antikörper (Serum, der wässrige Bestandteil des Blutes)			keine	keine	
Blutgruppe	A	B	A/B	0	

ABBILDUNG 47: ANTWORTBEISPIELE ZUM ITEM BLU05 ZU DEN TEILKOMPONENTEN MERII, GENETIK, SCHEMA (VGL. MUSTERLÖSUNG IM ANHANG 14.1.1)

In den Antwortbeispielen ist auffällig, dass die Konstruktion der Antikörper bei Blutgruppe A sowie die Zuordnung der Blutgruppe B in Spalte zwei der untersuchten Stichprobe weniger schwer fallen. Zur Konstruktion der Antigene für die Blutgruppe AB scheint es Probleme in der Konstruktion von „Kreis und Dreieck“ zu geben. Fälschlicherweise werden sie weggelassen anstatt beide gemeinsam aufzutragen. Im Hinblick auf die Konstruktion der Antikörper für Blutgruppe 0 liegen die Schwierigkeiten einerseits in der falschen Annahme, dass für Blutgruppe 0 ebenfalls keine Antikörper vorhanden sind oder andererseits in der gleichzeitigen Zuordnung der Symbole „Halbkreis“ und „Halbdreieck“ auf einem Antikörper statt dem parallelen Vorliegen beider Antikörper an sich.

9.5.3 TRANSLATION UND TRANSFORMATION AUS MEHREREN REPRÄSENTATIONEN (MERIII)

In den folgenden Beispielen werden Schwierigkeiten bei Aufgaben mit MERIII-Integration vorgestellt. Unabhängig davon, ob die Aufgaben anschließend der MERI-Integration oder der MERII-Integration zugeordnet wurden und damit auch das nach Modellprüfung gültige 2D-Modell angenommen wird, zeigen die Antworten fachdidaktisch interessante Muster und Schwierigkeiten in der Bearbeitung.

Die Testaufgabe See09 (Aufgabenkontext Ökologie, Repräsentationstyp Diagramm) stellt die insgesamt schwierigste Aufgabe in der vorliegenden Untersuchung dar ($M=2,702$, $SD=0,374$). Die Aufgabe wurde bei $n=106$ vorliegenden Antworten insgesamt nur mit knapp zehn Prozent richtig gelöst und zeigt in den weiteren deskriptiven und probabilistischen Statistiken Kennwerte im Grenzbereich auf. Aus fachdidaktischen Gründen wurde das Item See09 nicht aus dem Itempool entfernt³. Für die korrekte Antwort wird zunächst eine Informationsentnahme aus drei gegebenen Diagrammen zu unterschiedlichen Monaten im Jahresverlauf eines Sees gefordert. Im Anschluss daran erfolgt ein Repräsentationswechsel, indem die Information aus dem Text (die Algendichte von 5000 mL entspricht der Faulschlammdicke von 1m) in Verbindung mit den drei Diagrammen gebrachten werden muss. Nach erfolgter Umrechnung müssen die Werte der Faulschlammdicke in das gegebene Diagramm eingetragen werden, wodurch ein weiterer Repräsentationswechsel erforderlich ist. Im Folgenden werden markante Fehler und Schwierigkeiten in der Beantwortung dieser Testaufgabe vorgestellt (vgl. ABBILDUNG 48, ABBILDUNG 49).

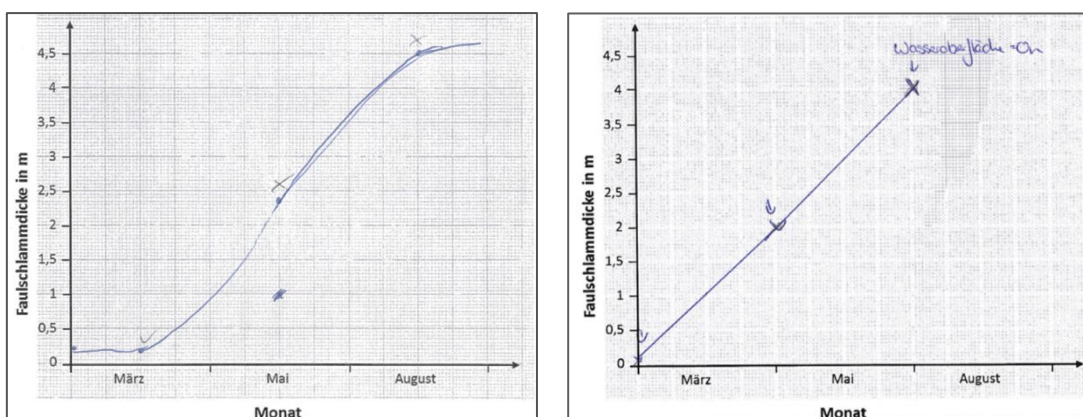


ABBILDUNG 48: ANTWORTBEISPIELE ZUM ITEM SEE09 - SKALIERUNG DER ZIELVARIABLE (MERIII, ÖKOLOGIE, DIAGRAMM)

³ Kennwerte zu Item See09: $I\sigma=2,702$ ($SE=0,374$), $wMNSQ=1,03$ ($CI=[0,42, 1,58]$), $T=0,2$, $Trennschärfe=0,02$. Es sei an dieser Stelle anzumerken, dass die Analysen nach Entfernung des Items aus dem Itempool zu keinen statistisch signifikant besseren Ergebnissen führten oder die Entfernung einen Einfluss auf die dargestellten Ergebnisse hat.

Wie die ABBILDUNG 48 zeigt, scheint die Wahl des Diagrammtyps sowie die Skalierung der Variable Faulschlammdicke [in m] den Studierenden Schwierigkeiten bereitet zu haben. So wurde die „Zeit“ (dargestellt in Monaten) als eine scheinbar metrisch skalierte Variable angenommen. Dadurch wurde ein fälschlicherweise kontinuierlicher Verlauf der Faulschlammdicke über die Monate als Liniendiagramm dargestellt. In anderen Fällen wurde ein ebenfalls kontinuierlicher Verlauf interpretiert, dieser jedoch in abnehmender Richtung, sodass zu Beginn der jeweiligen Monate die gesuchten Werte eingetragen werden konnten (vgl. ABBILDUNG 49). Ein Grund für diese nicht korrekte Interpretation kann im Repräsentationswechsel zwischen den drei gegebenen Diagrammen und der Textinformation „Dicke des Faulschlamm an der Wasseroberfläche“ gefunden werden. Auffällig dabei ist, dass in diesem Fall der Kurvenverlauf der drei gegebenen Diagramme auf das neu zu konstruierende Diagramm übertragen wurde.

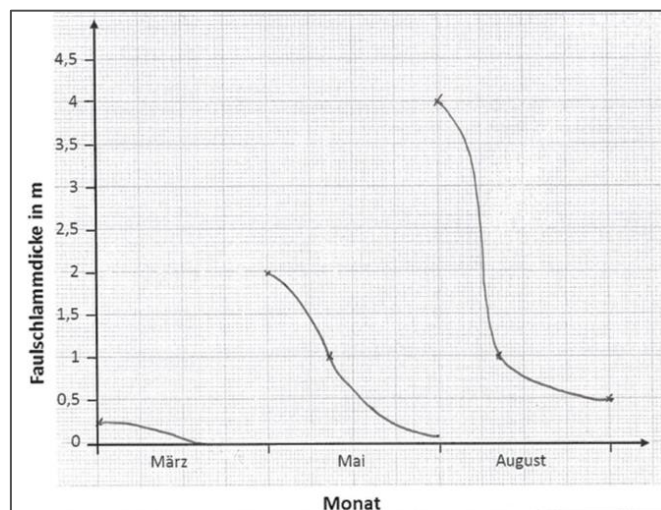


ABBILDUNG 49: ANTWORTBEISPIEL ZUM ITEM SEE09 - VERLAUF DER ZIELVARIABLE

Die Testaufgabe Erb05 (Aufgabenkontext Genetik, Repräsentationstyp Schema) beschreibt ein insgesamt tendenziell eher leichtes Item ($M=-0,313$, $SD=0,204$), das zu 57,8% richtig gelöst wurde ($n=109$ vorliegende Antworten). Für die korrekte Lösung wurde das Eintragen der fehlenden Anzahl der Geschlechtschromosomen in den Keim- und Körperzellen (F1) bei Hühnern gefordert. Antwortbeispiele zu dieser Aufgabe sind in ABBILDUNG 50 dargestellt.

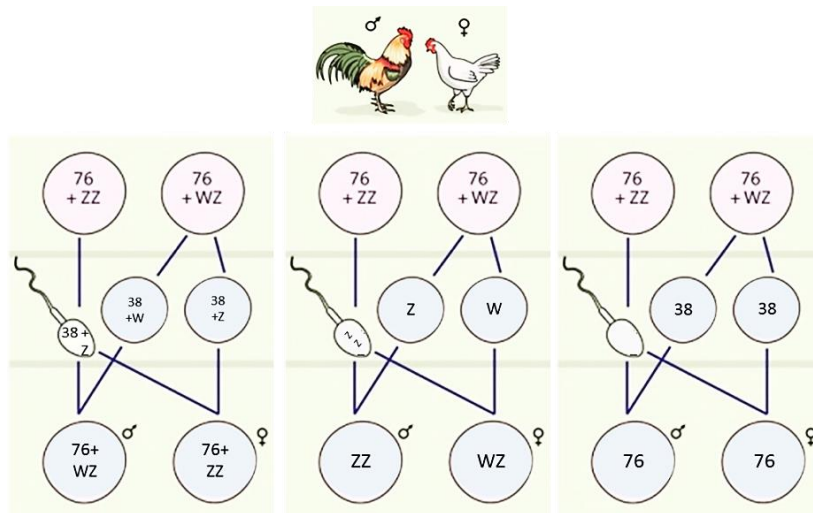


ABBILDUNG 50: ANTWORTBEISPIELE (A, B, C) ZUM ITEM ERB05 (MERIII, GENETIK, SCHEMA, VGL. MUSTERLÖSUNG ANHANG 14.1.1, S. 267)

Wie die ABBILDUNG 50 zeigt, haben Studierende scheinbar Schwierigkeiten in der richtigen Zuordnung der Gameten. Dabei folgt die Zuordnung gemäß der scheinbaren Richtung im Schema selbst, sodass im männlichen Organismus häufig $76+WZ$ und im weiblichen Organismus $76+ZZ$ eingetragen wurde. In einigen Fällen konnten die Gameten dem gegebenen Schema richtig zugeordnet werden, es fehlte jedoch die korrekte Nennung und anschließende Zuordnung der Chromosomenanzahl in den Keim- und Körperzellen (F1). Dieser Fall wurde ebenfalls umgekehrt beobachtet, das heißt die Anzahl konnte wiedergegeben werden, während die Nennung und Zuordnung der Chromosomenzahl ausblieb.

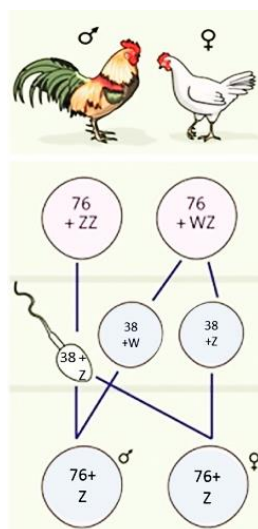


ABBILDUNG 51: ANTWORTBEISPIEL ZUM ITEM ERB05

In anderen Fällen fehlte die Übertragung der im Text und aus dem Schema zu entnehmenden Information, dass Keimzellen haploid und die Körperzellen der F1-Generation wieder als diploider Chromosomensatz vorliegen (vgl. ABBILDUNG 51).

V DISKUSSION

Im Folgenden werden die Ergebnisse der Hauptstudie hypothesenbezogen diskutiert und unter Rückbezug auf die theoretischen Annahmen interpretiert. Divergierende Befunde bezüglich der erhobenen Vorstudie werden ebenfalls erörtert. Für die Diskussion werden Bezüge zwischen den Ergebnissen der durchgeführten Erhebung und den vorgestellten theoretischen Modellen hergestellt. Hierfür werden in Kapitel 10 zunächst die methodischen Ansätze der Studie zusammenfassend dargelegt und die psychometrische Qualität des Kompetenztests zur Erfassung von biologiespezifischem Bildverständnis beim Umgang mit MER bewertet. In einem zweiten Abschnitt folgt schließlich die inhaltliche Diskussion, die sich auf die Struktur des entwickelten Kompetenzmodells und schwierigkeitsgenerierende Merkmale bezieht. Die Kernaussagen der Arbeit werden entsprechend in Kapitel 11 zusammengefasst. Abschließend werden in Kapitel 12 die fachdidaktische Relevanz sowie Implikationen für die Unterrichtspraxis und Forschung dargelegt.

10 METHODISCHE DISKUSSION

Vor dem Hintergrund der zugrundeliegenden Fragestellungen und der empirisch gewonnenen Daten erfolgt zunächst eine Zusammenfassung und kritische Beleuchtung des methodischen Vorgehens. Hierbei wird auf die Testentwicklung und die analysierte Stichprobe eingegangen (Kapitel 10.1). Im weiteren Verlauf wird die Wahl der Methoden zur Skalierung der durchgeführten Rasch-Analysen und für das eingesetzte Testdesign erörtert und ggf. mögliche Alternativen aufgezeigt (Kapitel 10.2). Die einzelnen Aspekte sowie Limitierungen der vorliegenden Arbeit werden kritisch reflektiert.

10.1 TESTENTWICKLUNG, STICHPROBE UND VALIDIERUNG

Für die empirische Überprüfung des postulierten Kompetenzstrukturmodells wurden insgesamt 210 Testaufgaben entwickelt und im Rahmen der Pilotierung auf ihre psychometrische Qualität überprüft (s. Kapitel 6). Aufgrund ihrer statistischen Kennwerte, ungünstiger Itemformulierungen und der inhaltlichen Eingrenzung auf ausgewählte Kontexte verblieben zunächst 186 Testaufgaben, wovon 95 für weiterführende Untersuchungen und abschließend 55 Items für die engere Auswahl zur Durchführung der Hauptstudie analysiert wurden (s. Kapitel 6.5). Daneben liegt ein ausführliches Kategoriensystem zur Kodierung offener Aufgaben bei der Lösung von Aufgaben zur Text-Bild-Integration vor (s. Kapitel 6.1). Der Pilotierung ist es außerdem zuzurechnen, dass der Leistungstest insgesamt eine angemessene Schwierigkeit für

die verhältnismäßig heterogene Stichprobe aufweist. Alle ausgewählten Testaufgaben der Hauptstudie (mit der Ausnahme des Items See09) genügen den statistischen Anforderungen und bestätigen die Wichtigkeit von Pilotierungen bei der Durchführung quantitativer Untersuchungen.

Mit der Überprüfung von Zusammenhängen zwischen den erhobenen Schulnoten und den Personenparametern kann gezeigt werden, ob der vorliegende Test biologiespezifische Kompetenzen erfasst. Damit kann die Konstruktvalidität geprüft werden. Aufgrund der signifikant höheren Korrelation zwischen der Biologienote im Vergleich zur Mathematiknote und den Personenparametern kann davon ausgegangen werden, dass vor allem biologische Kompetenzen diagnostiziert wurden. Dabei kann die nur schwache Korrelation als Hinweis darauf interpretiert werden, dass der vorliegende Kompetenztest vor allem Wissen erfasst, das so nicht üblicherweise abgefragt wird und bei der Notengebung eine Rolle spielt, oder es sich vorrangig um repräsentationsspezifisches Wissen handelt. Zusätzlich kann somit davon ausgegangen werden, dass die vorwissensunabhängige Konstruktion der Testaufgaben weitgehend gelungen ist. Insgesamt ist die Verwendung von Schulnoten als Prädiktor für biologiespezifisches Wissen nicht ganz unproblematisch, da aufgrund der Selbstauskunft Verzerrungen durch die soziale Erwünschtheit vorliegen können. Andererseits ist die Vergleichbarkeit der erfragten Noten schwierig zu interpretieren, da hinter ihnen unterschiedliche Bewertungssysteme und Einflussfaktoren stehen können (vgl. Schrader & Helmke, 2002). Alternative Referenzen wären in diesem Zusammenhang wünschenswert, gerade wenn es um Studienanfängerinnen und Studienanfänger geht, in denen vor allem die Eingangsvoraussetzungen für ein naturwissenschaftliches Studium interessant sind.

Die im Rahmen der Testentwicklung erstellten Items als operationalisierte Teile des Kompetenzkonstrukts dienen der Untersuchung und Überprüfung von Validität (vgl. Ropohl, Walpuski, & Sumfleth, 2014; Mayer & Wellnitz, 2014; Terzer et al., 2013; Schecker & Parchmann, 2006). Hierfür wurden die kognitiven Prozesse bei der MER-Integration zur Lösung der Testaufgaben definiert (vgl. Hartig, Frey, & Jude, 2012). Der Nachweis, dass die empirische Aufgabenschwierigkeit auf a priori definierte Merkmale zurückgeführt werden kann, dient als Bestätigung der Validität des Testinstruments (Hartig & Jude, 2007). In diesem Zusammenhang sind weitere Untersuchungen hinsichtlich formaler Merkmale wie z.B. die Textlänge und Teststruktur oder das Sprachniveau anzustreben (z.B. Rumann et al., 2010). Durch das Vorliegen einzelner Testaufgaben in dieser Untersuchung anstatt von so genannten Testlet-Aufgaben kann jedoch vorerst davon ausgegangen werden, dass der Split-Attention-Effekt reduziert und damit der Beitrag zum „extraneous load“ vor dem Hintergrund der *Cognitive Load Theory*

(s. Kapitel 2.2 und Kapitel 2.3) auf einem insgesamt niedrigen Niveau liegt und das Testergebnis nicht besonders stark beeinflusste (vgl. z.B. Ayres & Sweller, 2005; Sweller et al., 1998; Saß, 2012). Zur Messung von Fähigkeiten im Bereich der MER-Integration enthalten die Aufgabenstämme stets einen Text und ein oder mehrere Bilder im Aufgabenstamm. Die Messung der jeweiligen Teilkompetenz erfolgt somit über das jeweilige Testitem, zu dessen Lösung Informationen aus Text und Bild oder Text und Bildern integriert werden mussten. Die Ergebnisse aus der qualitativen Fehleranalyse (s. Kapitel 9.5) zeigen jedoch, dass einige Studierende Items zur *Konstruktionsnahen Integration (MERII)* oder zur *Translation oder Transformation aus mehr als zwei Repräsentationen (MERIII)* auch dann versucht haben zu lösen, wenn sie den zugehörigen Text nicht verstanden oder nicht aktiv integriert haben. Somit kann nicht ausgeschlossen werden, dass Probleme bei der MER-Integration auch zu Schwierigkeiten bei den Aufgaben zur *Informationsentnahmenahen Integration (MERI)* bzw. zur *MERIII-Integration* führen können. Die Messung der Fähigkeiten bei der *Konstruktionsnahen Integration (MERII)* führte aufgrund des verwendeten offenen Antwortformats zu Problemen in der Auswertung und damit zu Einbußen in der psychometrischen Qualität. Gleichzeitig steht aus fachdidaktischen Gründen die Messung einer solchen konstruktionsnahen Integrationskompetenz im Vordergrund. Eine mögliche Alternative für eine bessere Auswertungsobjektivität und eine damit verbundene Reliabilitätssteigerung könnten Multiple-Choice-Formate sein, die unterschiedliche Bilder als Attraktor und Distraktoren anbieten und die Testpersonen daraus die korrekte Antwort auswählen müssen. Saß (2012) konnte zeigen, dass solche Aufgaben im Vergleich zu Aufgaben aus rein textlich angebotenen Antwortalternativen signifikant häufiger gelöst werden. Dadurch sind hingegen Einbußen bei der Messung des integrativen Prozesses im Rahmen der Konstruktion zu erwarten, was es zunächst separat zu untersuchen gilt. Positiv wären jedoch eine zu erwartende kürzere Beantwortungszeit von Konstruktionsaufgaben sowie das Testen am Computer im Vergleich zu Paper-and-Pencil-Tests.

Für die Modellierung von Repräsentationskompetenz wurde versucht, Studierende im ersten oder zweiten Fachsemester sowie aus verschiedenen Studiengängen zu gleichen Teilen zu berücksichtigen (s. Kapitel 4.1 und Kapitel 7.1). Eine gleichmäßige Verteilung über die Studiengänge hinweg war dennoch zum Teil schwer zu realisieren, da die unterschiedliche Anzahl von Studierenden in den zur Verfügung stehenden Vorlesungen nicht immer eine homogene Zusammensetzung garantierte. Gleichzeitig war vor Durchführung der Erhebung nicht bei allen Veranstaltungen bekannt, welche Studiengänge in der Stichprobe mit welcher Häufigkeit vertreten sein werden. Die leichte Ungleichverteilung kann im Rahmen der vorliegenden Un-

tersuchung insgesamt vernachlässigt werden, da nicht davon auszugehen ist, dass die Variation der untersuchten Kompetenz im Verhältnis zur Stichprobe eingeschränkt ist (vgl. Rost, 2004, S. 72 f.; Schmiemann, 2010, S. 146). Die in diesem Zusammenhang stehende, geringe Varianzaufklärung kann dies ebenfalls bestätigen (s. Kapitel 9.4.2).

10.2 RASCH-SKALIERUNG UND TESTDESIGN

Im Rahmen der Modellprüfung ist eine durch den Test ausreichend genaue Schätzung der Itemschwierigkeiten und Personenparameter sowie eine hinreichende Reliabilität für die Analysen auf Populationsebene entscheidend (Terzer et al., 2013). Durch das Spannungsverhältnis aus verschiedenen Nutzungsperspektiven sind unterschiedliche Anforderungen an die verwendeten Testinstrumente und das Testdesign erforderlich (vgl. Fleischer et al., 2013; Klieme & Rakoczy, 2008). Diese sowie die Art der bereitgestellten Ergebnisse als Resultat der Testung sind auf ihre psychometrische Qualität zu überprüfen. Begründungen für die methodische Wahl sowie Limitierungen im Hinblick auf das Testdesign und die Rasch-Skalierung in Zusammenhang mit dem Antwortformat werden im Folgenden näher vorgestellt und erörtert.

In der vorliegenden Arbeit wurden insgesamt 210 Items für den Kompetenztest konstruiert und anschließend pilotiert. Aufgrund verschiedener Überlegungen wie Testlänge, Testheftdesign und Antwortformat (s. Kapitel 7.2) wurden schließlich 36 Testitems in der Hauptstudie eingesetzt. In der probabilistischen Testtheorie wird die Passung der Items zum Raschmodell über die Fit-Statistik überprüft (vgl. Bond & Fox, 2007). Da sich die Zielstellung der vorliegenden Studie auf die Populationsebene bezieht, ist die Reliabilität individueller Messergebnisse insgesamt von nachgeordneter Bedeutung (vgl. Terzer et al., 2013). Die bereits in der Pilotierung gute Item Separation Reliabilität konnte in der Hauptstudie weiter verbessert werden, was als Hinweis für eine präzise Schätzung der Itemschwierigkeiten interpretiert werden kann. Diese Steigerung fiel erwartungsgemäß aus, da weniger Items bei einer verhältnismäßig gleich großen Stichprobe eingesetzt wurden. Die klassischen Trennschärfen über die wMNSQ-Werte liegen sowohl für die Pilotierung als auch für die Hauptstudie in einem sehr guten Bereich.

Als problematisch ist hingegen die EAP/PV Person Separation Reliabilität anzusehen, die sich auf die Messgenauigkeit von Items bezieht und für eindimensionale sowie mehrdimensionale Konstrukte erfasst wird (vgl. Terzer et al., 2013). Dabei gilt es zu berücksichtigen, dass bei Vorliegen einer Mehrdimensionalität des Konstrukts die Reliabilität für jede angenommene Dimension separat betrachtet wird.

Zusammenfassend konnte die EAP/PV Person Separation Reliabilität in der Hauptstudie gegenüber der Vorstudie nicht ausgebaut werden. Bei der Interpretation der empirischen Daten ist anzumerken, dass jede Testperson insgesamt nur maximal neun Testaufgaben bearbeitet hat, d.h. für jede Dimension bzw. Teilkompetenz deutlich weniger als in der Pilotierung. Dies führt zu individuellen Messungenauigkeiten, die sich auf die Reliabilitäten der modellierten Skalen auswirken können (vgl. Adams, 2005). Der Effekt, dass gering bearbeitete Itemzahlen zu Einbußen bei der Reliabilität führen können, wird bereits in anderen Studien berichtet und reflektiert (z.B. Ropohl, 2010; Schmiemann, 2010; Terzer, 2012; Wellnitz, 2012; Ziepprecht, 2016). In diesem Zusammenhang könnte auch die Verwendung des gewählten Testdesigns eine Rolle spielen. Zur Operationalisierung des postulierten Kompetenzstrukturmodells wurden die konstruierten Testaufgaben in einem Youden-Square-Design zur Beantwortung vorgelegt (s. Kapitel 4.3). Als besondere Form eines *Balanced Incomplete Block Design (BIBD)* werden dabei Positionseffekte kontrolliert (vgl. Frey, Hartig, & Rupp, 2009). Problematisch ist jedoch, dass Carry-Over-Effekte nicht kontrolliert werden und damit das Testergebnis beeinflussen können. Simulationsstudien zeigen, dass die Modellfitwerte immer auch von den gebildeten Clustern, d.h. der Itemblöcke, sowie dem verwendeten Testdesign abhängen (z.B. Brandt, 2012).

Für die verhältnismäßig geringen Reliabilitäten der ein- und mehrdimensionalen Skalen lassen sich weitere Gründe anführen. So ist anzumerken, dass die Reliabilität im Zusammenhang mit der Varianz steht. Vergleicht man die Testkennwerte der ein- und mehrdimensionalen Rasch-Skalierungen (vgl. TABELLE 14 für die Pilotierung sowie TABELLE 29 für die Hauptstudie) stellt man fest, dass die Streuung der Kennwerte insgesamt eher gering ausfällt und die geringe Merkmalsstreuung wiederum eine hohe Korrelation unwahrscheinlich werden lässt. Die Varianz der Personenfähigkeiten lag in der Hauptstudie mit Ausnahme der Genetik-Skala für alle Modelle unterhalb von Eins, was als Erklärungsansatz für die geringen Reliabilitäten herangezogen werden kann (vgl. Adams, 2005). In diesem Fall wäre eine entsprechende Testlänge unter Verwendung homogener Items erforderlich (z.B. Amelang & Zielinski, 2002; Terzer et al., 2013). Damit verbunden ist auch eine Erhöhung der Itemanzahl pro Dimension. Vor diesem Hintergrund sollten auch die Testitems der MERIII-Integration reflektiert werden und eine gezielte Variation und Verteilung je nach Output (Text oder Bild) erfolgen. Die Ergebnisse liefern erste Hinweise darauf, dass es sich um eine separate Fähigkeit handelt, wenn der Repräsentationswechsel von mehreren Repräsentationen erfolgen muss (vgl. TABELLE 30). Da es sich in der vorliegenden Arbeit um einen Kompetenztest handelt und zur Überprüfung sowie Validierung sehr unterschiedliche Testaufgaben in verschiedenen Aufgabekontexten eingesetzt wurden, kann von einer Heterogenität der Testitems ausgegangen werden. Wohingegen homogene Tests fast immer eine hohe innere Konsistenz aufweisen,

da sie meist ähnliche oder gleiche Items enthalten, trifft dies auf Testaufgaben in Wissens-tests nicht zu (vgl. Bühner, 2011). Entsprechend ist eine hohe Reliabilität ebenfalls nicht zu erwarten, was durch die maximale Anzahl von 36 eingesetzten Testaufgaben zusätzlich beeinflusst wird. Darauf folgt, dass eine Erhöhung der Itemzahl anzustreben ist.

Als weiterer Grund ist anzuführen, dass in der Hauptstudie mehr als 94% der eingesetzten Items im offenen Antwortformat vorliegen (vgl. TABELLE 3, s. Kapitel 4.2). Testinstrumente mit offenem Antwortformat weisen oftmals geringere Personenreliabilitäten auf als Untersuchungen, die zu einem überwiegenden Teil aus geschlossenen Testaufgaben bestehen (z.B. Stecher & Klein, 1997). In diesem Zusammenhang könnte auch die Motivation der Testpersonen für die Beantwortung offener, z.T. zusätzlich fachfremder, Aufgaben stehen. Rauch & Hartig (2010) konnten zeigen, dass Probanden eher Multiple-Choice präferieren, wenn sie die Wahl haben. Gleichzeitig könnte dabei auch die Bereitschaft Zeit in einen derartigen Test zu investieren, eine Rolle beim Antwortverhalten spielen. Dadurch könnte es sein, dass die tatsächliche Testleistung der Probanden nicht vollständig abgebildet werden konnte und es zu Verzerrungen im Testergebnis gekommen ist. Für die anschließende Rasch-Skalierung wurden die Items dichotom codiert, was zu einem Informationsverlust infolge der Kategorisierung sowie zu Einbußen in der trennscharfen Bewertung geführt haben könnte. In diesem Zusammenhang steht auch eine bei der Kodierung resultierende mangelnde Auswertungsobjektivität, die zu einer niedrigeren Testmodellpassung führen kann (z.B. Spoden, Fleischer, & Leutner, 2014). In der Studie konnte nachgewiesen werden, dass die Auswertungsobjektivität einen Einfluss auf die Modellpassung hat und eine geringe Objektivität Abweichungen in den Personenparametern begründen können. So wurden in der vorliegenden Studie zwar hohe Interrater-Reliabilitätswerte erreicht, dennoch sind solche Effekte nicht auszuschließen. Ein weiterer Grund für die niedrigen Reliabilitäten könnte in der Verletzung der lokalen stochastischen Unabhängigkeit als Voraussetzung für IRT-Modelle liegen. Die Art der Beantwortung einer Aufgabe sollte unabhängig davon sein, wie andere Aufgaben im selben Test beantwortet werden (vgl. Hartig & Frey, 2013). Gerade bei Aufgaben mit gemeinsamen Aufgabenstämmen, wie sie in der Pilotierung vorliegen, oder wenn Messwiederholungen auftreten, wird die zuverlässige Schätzung erschwert und kann nicht immer gewährleistet werden. Eine weiterführende Analyse unter Kontrolle des Antwortformats wäre somit wünschenswert. Zusätzliche kann eine polytome Codierung der offenen Aufgaben zu mehr Aufklärung führen, weil die Antworten nicht nur richtig oder falsche Antworten (Full- und No-Credit-Verfahren), sondern auch teilweise richtige Lösungen zulassen (vgl. Hammann & Jördens, 2014). In diesem Zusammenhang sind Partial Credit Modelle heranzuziehen, die zusätzlich als Hinweis für mögliche Niveaustufen dienen können (vgl. Rost, 2004; Embretson & Reise, 2000).

Die niedrigen Reliabilitäten der Gesamtskala und der Einzelskalen könnten auch ein Hinweis darauf sein, dass ein einheitliches, integrations- und kontextunabhängiges Kompetenzkonstrukt vorliegt, die ermittelten informationstheoretischen Maße weisen jedoch auf ein mehrdimensionales Kompetenzmodell hin (s. Kapitel 8.2.2, TABELLE 30). Dennoch muss angenommen werden, dass durch die Items der verschiedenen Einzelskalen weitere, nicht modellierte Faktoren zusätzlich erfasst wurden, da nicht davon auszugehen ist, dass das gesamte Wissensspektrum durch eine einzige Testform abzufragen ist (vgl. Schmiemann, 2010). Zusätzlich werden für dieselbe Datenlage nur die Maße AIC, BIC und CAIC für die konkurrierenden Modelle herangezogen, da Maße und Tests zur Einschätzung der absoluten Passung der Daten auf das Modell aktuell noch fehlen (Hartig & Frey, 2013). Rost (2004) weist hingegen darauf hin, dass Modellgeltungstests mit den Kriterien AIC und BIC bereits ab zwei Dimensionen problematisch sind, da beide die Parameterzahl berücksichtigen. Daneben müsste auch der Kontext, in dem die Messung durchgeführt wurde, oder das Ziel der Messung (z.B. High-Stake-Tests vs. Low-Stake-Tests) berücksichtigt werden, indem zum Beispiel Mischverteilungsmodelle verwendet werden (z.B. Rost & von Davier, 1995).

Weitere Arten von Messfehlern, die zufällig oder spezifisch im Rahmen der Untersuchung aufgetreten sind, können zusätzlich eine Rolle spielen. So können u.a. Begriffe wie die hier verwendeten Operatoren einen Faktor darstellen, der die Reliabilität beeinflusst (Bühner, 2011). In diesem Zusammenhang sollte eine Post-Hoc-Überprüfung der Operatoren durchgeführt werden. Vor diesem Hintergrund sind die geringen Reliabilitätswerte im Bereich bis 0,42 jedoch insgesamt erwartungsgemäß. Trotzdem entspricht dies nicht den Grenzwerten, die im Rahmen der klassischen Testtheorie zur Interpretation von Cronbachs Alpha herangezogen werden (vgl. Rost, 2004). Eine Erhöhung der Reliabilität ist aus psychometrischen Gründen damit anzustreben. Im technischen Bericht der PISA-Studie 2012 werden für die beiden eindimensionalen Skalen „*Digital reading*“ und „*Computer-based mathematics*“ ebenfalls sehr geringe Reliabilitäten (0,22 und 0,20) dokumentiert (vgl. OECD, 2014, S. 230). Bei einer mehrdimensionalen Skalierung unter Berücksichtigung von Hintergrundmodellen wie z.B. das Testdesign, steigen die Reliabilitäten der beiden Skalen auf über 0,91 (vgl. OECD, 2014, S. 233). Vor diesem Hintergrund sind weitere statistische Verfahren wie z.B. die Spearman-Brown-Korrektur (z.B. Schermelleh-Engel & Werner, 2012), die das vorliegende Testdesign, die heterogene Stichprobe sowie das Antwortformat berücksichtigen, zur Überprüfung der Reliabilitäten heranzuziehen. Die Qualität des Tests kann zu diesem Zeitpunkt jedoch über die sukzessive Entwicklung, ausgehend von einer theoretischen Fundierung bis zu etablierten Validierungsschritten, argumentiert werden (Terzer et al., 2013). Dennoch bleibt anzumerken, dass die geringen Schätzer für die Reliabilität die Interpretierbarkeit der berichteten Ergebnisse teilweise einschränken.

Zu den erhobenen Kontrollvariablen Biologie- und Mathematiknote in der vorliegenden Arbeit hätte das für die Aufgaben erforderliche biologische Fachwissen erhoben werden können. Im Kompetenztest wurden Testaufgaben zu unterschiedlichen Aufgabenkontexten der Biologie entwickelt, die im Unterricht der Sekundarstufe II und dem Abitur eine Rolle spielen. Durch diese thematische Vielfalt wurde mithilfe des Kompetenztests das Repräsentationswissen über die unterschiedlichen biologischen Inhalte erhoben und gleichzeitig untersucht, ob der Aufgabenkontext ein Prädiktor für die Schwierigkeit der Items darstellt. Die möglichst vorwissensunabhängige Gestaltung der Testaufgaben diene zur Kontrolle des notwendigen Fachwissens, dennoch hätte ein umfangreicher Fachwissenstest eingesetzt werden können, um diesen Einfluss auch statistisch zu überprüfen. Ein solcher Test hätte jedoch zu einer übermäßigen Erhöhung der Testzeit geführt, trotz der möglichst unabhängig vom Fachwissen gestalteten Aufgaben ist jedoch davon auszugehen, dass das Fachwissen dennoch eine Rolle bei der Beantwortung der Testaufgaben spielt.

Angesichts der niedrigen Reliabilitäten der Gesamtskala sowie der Einzelskalierungen der verschiedenen Teilkompetenzen sollten die Ergebnisse vorsichtig interpretiert werden. Aufgrund der Kategorisierung von Itemmerkmalen, die mittels regressionsanalytischer Verfahren zur Vorhersage von Itemschwierigkeiten herangezogen wird (vgl. Gorin & Embretson, 2006), können Aussagen nur über korrelative Zusammenhänge getroffen werden, die jedoch keine kausalen Schlussfolgerungen erlauben (Saß, 2012). Die IRT-Modelle identifizieren zwar verschiedene Kompetenzausprägungen der untersuchten Personen und können erste Hinweise auf Kompetenzniveaus abbilden, allerdings liefern sie uns keinen Hinweis darauf, wie wir solche Kompetenzpunkte erhöhen können (vgl. Schnotz et al., 2011). In diesem Zusammenhang sind experimentelle Zugänge wünschenswert (z.B. Haladyna & Downing, 2004), die eine systematische Manipulation von Aufgabenmerkmalen ermöglichen und dabei die kognitiven Prozesse unter Berücksichtigung von Strategien und dem vorhandenen Wissen abbilden. Dadurch können Merkmale identifiziert werden, die diese Prozesse beeinflussen und deren Wirkung auf die Bearbeitung von Testaufgaben überprüft werden kann (vgl. Saß, 2012, S. 21). Dennoch liefert die vorliegende Auswertung aufgrund ihrer Detailliertheit besonders nützliche Erkenntnisse für weiterführende Untersuchungen hinsichtlich Kompetenzniveau Modelle sowie neue empirische Evidenz was die Entwicklung und Validierung von biologiespezifischen Testaufgaben betrifft.

11 INHALTLICHE DISKUSSION

In der folgenden inhaltlichen Diskussion werden die Ergebnisse zusammenfassend dargelegt und Erklärungsansätze für die beschriebenen Befunde erläutert. Dabei werden die Ergebnisse insbesondere in Bezug auf die Struktur von Repräsentationskompetenz als *Informationsentnahmenahne Integration (MERI)* und *Konstruktionsnahe Integration (MERII)* sowie ihrer Relation zueinander erörtert. Die Diskussion wird in Hinblick auf die formulierten Ziele, Forschungsfragen und Hypothesen in Abschnitt III geführt. In einem ersten Schritt wird zunächst die postulierte Kompetenzstruktur vor dem Hintergrund unterschiedlicher Dimensionierungen erläutert (s. Kapitel 11.1). Anschließend werden die Befunde bezüglich Leistungsvarianz sowie Zusammenhänge zu weiteren Aufgabenmerkmalen diskutiert (s. Kapitel 11.2 und Kapitel 11.3).

11.1 MODELLVERGLEICH UND MODELLSTRUKTUR

Auf Basis von theoretischen Annahmen wurde ein Kompetenzstrukturmodell zum Umgang mit multiplen externen Repräsentationen (MER) entwickelt und auf seine empirische Gültigkeit überprüft. Die Passung der verschiedenen Modelle zu den ermittelten empirischen Daten wurde mithilfe der probabilistischen Testtheorie überprüft. Beim Modellvergleich zeigt sich aufgrund der informationstheoretischen Maße AIC, BIC und CAIC für das zweidimensionale Modell eine bessere Passung als für das eindimensionale Modell und die weiteren Alternativmodelle (s. Kapitel 8.1.3). Die Berechnung dieser informationstheoretischen Maße führte im Vergleich zur Final Deviance zu heterogenen Befunden (vgl. TABELLE 30). Während unter Bezugnahme auf die Final Deviance das sechsdimensionale within-Modell am besten zu den Daten passt, zeigt sich über alle herangezogenen Maße hinweg, dass das zweidimensionale Modell zur MER-Integration die beste Passung erreicht, wodurch die Hypothesen $H_{1,0}$, $H_{1,1}$, $H_{1,2}$ und $H_{1,3}$ verworfen werden. Gemäß diesem Modell werden zwei Teilkompetenzen unterschieden: Lernende müssen über bestimmte Fähigkeiten zur *Informationsentnahmenahnen Integration* verfügen sowie über andere Fähigkeiten, die eine *Konstruktionsnahe Integration (MERII)* abbilden. Die hohen Korrelationen zwischen den beiden Teilkompetenzen verweisen erwartungskonform auf Gemeinsamkeiten in den Anforderungen der Kompetenzanteile, sind aber deutlich unter dem Referenzwert, weshalb von einem mehrdimensionalen Konstrukt auszugehen ist (vgl. TABELLE 31). Die durch das Kompetenzstrukturmodell postulierte Struktur (vgl. ABBILDUNG 7) kann damit nur zum Teil durch die empirischen Daten bestätigt werden. Unter Berücksichtigung der informationstheoretischen Maße und ihren nur geringen Abständen zueinander, sind aus den genannten Gründen ebenfalls alternative Kompetenzmodelle

denkbar wie beispielsweise die beiden dreidimensionalen Modelle *MER-Integration* mit MERI, MERII, MERIII bzw. das *Kontextmodell* aus Ökologie, Stoffwechsel, und Genetik. Im Hinblick auf den Repräsentationswechsel, der eine Integration aus mehr als zwei Repräsentationen erfordern, ist eine dreidimensionale Kompetenzstruktur theoretisch weiterhin anzunehmen. Die Ergebnisse liefern hierzu ebenfalls erste empirische Hinweise. So ist es eine besondere Schwierigkeit, wenn beispielsweise aus in der Aufgabenstellung vorgegebenen Diagrammen ein neues Diagramm transformiert werden muss, wodurch mehrere Integrationsprozesse nacheinander stattfinden (vgl. Item See09). In diesem Zusammenhang sind weitere Untersuchungen anzustreben, um diese Teilkomponente auch empirisch abzugrenzen. Für die beiden Teilkompetenzen MERI und MERII konnte gezeigt werden, dass diese die Schwierigkeit der Testitems beeinflussen, wodurch die Hypothese $H_{2.1}$ für die beiden Teilkomponenten MERI und MERII bestätigt wird. Die drei Teilkomponenten zur *MER-Integration* MERI, MERII, MERIII bzw. die zwei Teilkomponenten MERI und MERII sowie die drei unterschiedlichen *Aufgabenkontexte* Ökologie, Stoffwechsel, Genetik und die zwei *Repräsentationstypen* Diagramm und Schemazeichnung erklären zusammen 35% der Itemschwierigkeit, wobei der Haupteffekt mit einer Varianzaufklärung von etwa 16% in der *MER-Integration* liegt (s. Kapitel 9.3.3). Entsprechend ist davon auszugehen, dass die beiden Dimensionen *Aufgabenkontext* und *Repräsentationstyp* als Prädiktoren und damit als a priori definierte Aufgabenmerkmale in das Modell eingehen. Vor diesem Hintergrund können biologiespezifische Gestaltungsmerkmale und Konventionen innerhalb der einzelnen Themengebiete besser berücksichtigt werden. Die Befunde bezüglich dieser Aspekte werden in Kapitel 11.2. erläutert.

Insgesamt liegen damit erste empirische Evidenzen im Hinblick auf Aufgabenschwierigkeiten und mögliche Einflussfaktoren auf Kompetenzen zur Integration multipler externer Repräsentationen vor. Häufig wird erwartet, dass Testaufgaben je nach Art, Umfang und Verknüpfungen innerhalb der Aufgabenstellung sowie nach Art der Anforderungen an die informationsverarbeitenden Prozesse schwieriger oder leichter werden (z.B. Artelt et al., 2001; Kulgemeyer & Schecker, 2007; Klieme, 2000). Empirische Hinweise unter Berücksichtigung dieser Aufgabenmerkmale sind bisher jedoch nur wenig vertreten (vgl. Florian, Sandmann, & Schmiemann, 2014). So konnte Kauertz (2008) etwa 23% der Aufgabenschwierigkeit auf die in der Aufgabenstellung dargebotenen Inhaltsstrukturen aufklären. Schmiemann (2010) erreichte beim *konzeptuellen Wissen* und beim *begrifflich-sprachlichen Wissen* eine Varianzaufklärung von bis zu 55% der Aufgabenschwierigkeit. Erklärungsmodelle mit höheren Aufklärungsquoten wären wünschenswert, scheinen aber empirisch nur schwer zu erreichen sein, da aufgrund der Kompetenzmodellierung nur eine bestimmte Zahl von a priori Merkmalen zur

Erklärung zur Verfügung stehen (vgl. Schmiemann, 2010). Dessen ungeachtet zeigen Studien, dass der Kontext, in dem eine kognitive Leistung in einer spezifischen Testsituation zu erbringen ist, einen Einfluss auf die Qualität dieser Leistung haben kann (z.B. Blessing & Ross, 1996; Kirsh, 2009), aber es bisher nicht geklärt ist, welche Art von Kontext per se leistungsförderlich ist (Fleischer et al., 2014; vgl. auch Kapitel 9.3.2).

Dennoch stellt sich die Frage, welche Faktoren zur Erklärung der verbleibenden, nicht durch die Aufgabenmerkmale des Modells bestimmten Rests in der Varianz der Schwierigkeit herangezogen werden können. Zunächst sind weitere formale Aufgabenmerkmale, die einen Einfluss auf die Schwierigkeit einer Aufgabe haben können, zu kontrollieren. Häufig wird der Aufgabenkontext als Einflussfaktor genannt, wenn es um schwierigkeitsgenerierende Merkmale geht. Dieser ist in Zusammenhang mit repräsentationsspezifischem Wissen bisher jedoch kaum untersucht. Die vorliegenden Ergebnisse zeigen erste Hinweise, dass der inhaltliche Kontext für die Schwierigkeit einer Aufgabe eher eine untergeordnete Rolle spielt im Vergleich zur MER-Integration. Dies entspricht den Ergebnissen der Datenanalysen, die im Rahmen des PISA-Naturwissenschaftstests erhalten wurden (vgl. Prenzel et al., 2002). So zeigen die vorliegenden Ergebnisse, dass vor allem Aufgaben aus dem Bereich des Stoffwechsels besonders schwierig zu lösen waren. Eine mögliche Begründung könnte in der Wahrnehmung von Stoffwechselprozessen liegen. Ummels et al. (2015) berichten, dass Lernende häufig nicht in der Lage sind, die belebte mit der unbelebten Natur zu verknüpfen, d.h. sie haben Schwierigkeiten metabolische Prozesse und Prozesse der Energieerhaltung und Energieumwandlung auf die zelluläre Ebene zu übertragen. Das Schwierigkeitsbild ist insgesamt jedoch sehr heterogen, da es zusätzlich einen Unterschied macht, welche Art von Integration gefordert wird und welche Art von Repräsentationstyp herangezogen werden muss (vgl. Kapitel 9.3.2). So sind Diagrammaufgaben, die eine *Konstruktionsnahe Integration (MER-II)* erfordern, insgesamt am schwierigsten zu lösen, unabhängig vom Kontext Ökologie oder Stoffwechsel. Die Ergebnisse zeigen hingegen, dass ausschließlich der Aufgabenkontext Genetik, der wiederum nur aus konventionalisierten Schemata besteht, einen Einfluss auf die Aufgabenschwierigkeit hat, wodurch die Hypothese H_{2.2} für die Kontexte Stoffwechsel und Genetik und die Hypothese H_{2.3} verworfen werden. Dabei sind Genetikaufgaben vor allem für Mathematikstudierende besonders einfach zu lösen. Der Aufgabenkontext wurde in der vorliegenden Arbeit für die Aufgaben zwar systematisch variiert, dennoch kann die Tatsache, dass in der Hauptstudie für den Aufgabenkontext Stoffwechselphysiologie doppelt so viele Testaufgaben vorliegen, die Ergebnisse beeinflussen. Einen Erklärungsansatz für die statistisch nicht bedeutsamen Schwierigkeitsunterschiede in den drei untersuchten Aufgabenkon-

texten kann die vorwissensunabhängige Konstruktion der Testaufgaben liefern. Problematisch ist dabei, dass die Aufgaben gleichzeitig aufgrund der Berücksichtigung weiterer Teilkomponenten des Kompetenzmodells jedoch insgesamt sehr heterogen sind. Die signifikanten Effekte beim Kontext Genetik können aber als deutlicher Hinweis darauf interpretiert werden, dass hierzu weiterführende Untersuchungen notwendig sind.

Vor diesem Hintergrund bleibt festzuhalten, dass weitere Merkmale heranzuziehen sind, um mit den vorliegenden Vorhersagemodellen zusätzliche Erkenntnisse in der Aufklärung von Aufgabenschwierigkeiten beim Einsatz von biologiespezifischen Repräsentationen zu gewinnen. Gleichzeitig muss davon ausgegangen werden, dass eine perfekte Vorhersage der Aufgabenschwierigkeit durch Aufgabenmerkmale nicht zu erwarten ist (vgl. Prenzel et al., 2002).

11.2 DIFFERENZIERUNG UND AUSPRÄGUNG DER KOMPETENZEN

Zur fachdidaktischen Erforschung von Repräsentationskompetenz im naturwissenschaftlichen Kontext liegen aktuell nur wenige Studien vor. So werden zwar verschiedene Kompetenzteilbereiche unterschieden, aber nicht immer eine Dimensionalitätsprüfung durchgeführt (vgl. Ziepprecht, 2016). Lachmayer (2008) konnte für den Umgang mit Diagrammen die drei Komponenten Informationsentnahme, Konstruktion und Integration trennen und empirische Evidenz dafür vorlegen. In der Arbeit von von Kotzebue (2014) konnte die Text-Diagramm-Integration in die zwei Repräsentationszweige *Informationsnahe Integration* und *Konstruktionsnahe Integration* empirisch nachgewiesen werden. Ziepprecht (2016) untersuchte den Teilbereich *Biologische Informationen aus unterschiedlichen Repräsentationen erschließen* des Kompetenzbereichs Kommunikation an Schülerinnen und Schülern. Dabei konnte gezeigt werden, dass zwei Teilkomponenten im Rahmen der Informationsentnahme zu unterscheiden sind: *Biologische Informationen aus Texten erschließen* und *Biologische Informationen aus Bildern und Bild-Text-Kombinationen erschließen*. Die vorliegende Arbeit knüpft mit der Bestätigung des zweidimensionalen Modells für Repräsentationskompetenz, bestehend aus *Informationsentnahmenähe Integration (MERI)* und *Konstruktionsnahe Integration (MERII)*, an diese Arbeiten an und liefert Hinweise darauf, dass es bei Integrationsprozessen zwei zu unterscheidende kognitive Prozesse gibt. Die latenten Korrelationen liegen im mittleren Bereich. Damit kann davon ausgegangen werden, dass es sich um unterschiedliche latente Kompetenzkonstrukte handelt, gleichzeitig aber auch gemeinsame übergeordnete repräsentationsspezifische Fähigkeiten abgebildet werden. Daneben sind die beiden Teilkompetenzen für Studierende unterschiedlich schwierig und werden von unterschiedlichen Faktoren beeinflusst (s. Kapitel 9.5). Dies bestätigt Hypothese H_{3.1} für die Teilkomponenten MERI und

MERII. Die Differenzierung dieser zwei Teilkompetenzen, die sich im Rahmen der Modellprüfung ergab, deckt sich mit den Befunden bisheriger Studien, wenn es um den Output einer Aufgabenlösung geht. So stellt die Informationsentnahme aus einem Diagramm mit Output Text eine andere Fähigkeit dar als z.B. ein Diagramm zu konstruieren (z.B. Lachmayer, 2008; von Kotzebue & Nerdel, 2012). Typische Schwierigkeiten beim Umgang mit Diagrammen wie die falsche Wahl des Diagrammtyps, eine nicht angemessene Skalierung oder der Vertausch von abhängiger und unabhängiger Variable werden auch in der vorliegenden Arbeit gefunden (s. Kapitel 9.5 sowie u.a. von Kotzebue, Gerstl, & Nerdel, 2015; Delgado & Lucero, 2015; Lachmayer, Nerdel, & Prechtel, 2007; Baker, Corbett, & Koedinger, 2001). Während in diesen Studien, bezogen auf Diagramme, Konstruktionsaufgaben mit einer höheren Wahrscheinlichkeit gelöst werden, werden in dieser Untersuchung dazu konträre Ergebnisse gefunden (vgl. ABBILDUNG 43). Eine mögliche Erklärung kann die Anforderung einer vollständigen Diagrammkonstruktion sein, indem alle relevanten Schritte abgeleistet werden müssen. So können z.B. bereits Fehler in der richtigen Wahl des Diagrammtyps zu einer falschen Lösung insgesamt führen. Baker, Corbett, & Koedinger (2001) berichten, dass Aufgaben, die eine adäquate Wahl des Diagrammtyps erforderten, mit einer Wahrscheinlichkeit von weniger als 25% richtig gelöst wurden. Cheng & Gilbert (2015) sowie Hegarty (2011) führen an, dass die Schwierigkeit bei der Konstruktion von Diagrammen mit den zugrundeliegenden Charakteristika zusammenhängt. Von Kotzebue (2014) berichtet ebenfalls davon, dass bei der konstruktionsnahen Integration weniger das Fachwissen entscheidend ist, sondern das nötige Wissen über Konventionen. Gerade Diagramme zeichnen sich dadurch aus, dass sie eine komplexe Mischung aus konkreten und abstrakten Elementen, Konventionen sowie räumlichen Isomorphismen beinhalten.

Fehler in der Text-Bild-Integration können im Repräsentationswechsel liegen, wenn es darum geht, die richtigen Informationen aus dem Text mit dem gegebenen Bildtyp zu integrieren. Schnotz et al. (1999) zeigen, dass weniger erfolgreiche Lerner häufig daran scheitern, auf die notwendige und adäquate Information zurückgreifen und dadurch die bildlichen Informationen in einem wesentlich geringeren Ausmaß nutzen. Hochpöchler et al. (2013) führen ähnliche Befunde an und erklären, dass Lernende beim Text-Bild-Verstehen vor allem textgeleitet vorgehen und das Bild zunächst als weitere Informationsquelle nutzen. Dabei geht Schnotz (2005) davon aus, dass für das Text- als auch für das Bildverstehen der Rückgriff auf eigene konzeptuelle kognitive Schemata über Repräsentationswissen erfolgt. Empirische Evidenz liegt in diesem Zusammenhang für das Verstehen von Bildern vor (z.B. Kragten et al., 2013; Lowe, 1989). Zusätzlich belegen Studien, dass Lernende bessere Leistungen erbringen,

wenn sie bereits in unterschiedlichen Lerngelegenheiten mit Bildern oder Bild-Text-Kombinationen konfrontiert wurden. Vor diesem Hintergrund liefert die Untersuchung von Ziepprecht (2016) einen weiteren Beitrag, wenn es um das *Informationen erschließen aus einem Text, Bild oder einer Bild-Text-Kombination* geht. Hinsichtlich der Text-Bild-Integration berichtet Ziepprecht (2016), dass neben dem Textverstehen eine weitere Fähigkeit notwendig ist, die mit spezifischem Wissen über die Darstellungskonvention in biologischen Bildern einhergeht. Die vorliegende Arbeit bestätigt diese Befunde und zeigt anhand der untersuchten Bildtypen Diagramm und Schemazeichnung, dass unterschiedliche Kompetenzausprägungen in der MER-Integration und vor dem Hintergrund des eingesetzten *Repräsentationstyps* vorliegen. So konnten zwischen den beiden Typen Diagramm und Schemazeichnung bedeutsame Unterschiede in den Personenfähigkeiten diagnostiziert werden (s. Kapitel 9.4.1). Dies bestätigt Hypothese H_{3.3}. Ist den Lernenden beispielsweise nicht klar, dass die Energie auf den unterschiedlichen Ernährungsstufen durch die Balkenbreite dargestellt wird, fällt ihnen die Interpretation dieser Darstellungsform vermutlich eher schwer (vgl. Ziepprecht, 2016). Die Ergebnisse der qualitativen Fehleranalyse zeigen, dass in vielen Fällen nur eine Repräsentation herangezogen wird und die eigentliche Text-Bild-Integration ausbleibt (s. Kapitel 9.5). Dies ist beispielsweise der Fall, wenn Lernende entscheidende Informationen aus dem Text wie konkrete Werte oder Daten zur Diagrammkonstruktion entnehmen müssen, das Diagramm hingegen mit willkürlichen Abständen anschließend skaliert wird (vgl. Item See07, s. Anhang 14.1.1). Das mangelnde Wissen über fachspezifische Darstellungsformen und deren Konventionen sowie die Defizite im Bereich der Text-Bild-Integration können durch fehlende Lerngelegenheiten und Anwendungssituationen zur konkreten Unterscheidung verschiedener Repräsentationstypen erklärt werden (McElvany et al, 2010).

Für die Teilkomponente *Aufgabenkontext* liegen ebenfalls bedeutsame Unterschiede in den Personenfähigkeiten zwischen Stoffwechsel- und Genetik-Aufgaben sowie zwischen Ökologie- und Genetik-Aufgaben vor. Dadurch wird die Hypothese H_{3.2} nur für die Unterschiede zwischen Ökologie und Stoffwechsel verworfen. Entgegen den Erwartungen, dass insbesondere Genetikaufgaben für die Stichprobe schwierig zu lösen sind, indem das nötige Wissen über Konventionen erforderlich ist, wurden diese Aufgaben insgesamt am leichtesten gelöst. Die zugrundeliegende Konvention kann auch als eine mögliche Begründung für die ermittelten Unterschiede herangezogen werden. Nicht erwartungsgemäß waren die Ergebnisse zwischen Ökologie- und Stoffwechsel-Aufgaben, in denen keine statistisch bedeutsamen Unterschiede identifiziert werden konnten. Dabei sind Stoffwechsel-Aufgaben etwas schwieriger zu lösen als Ökologie-Aufgaben und wurden von der untersuchten Stichprobe am wenigsten gut gelöst. Die Ergebnisse können dahingehend interpretiert werden, dass Aufgaben

aus diesen beiden Kontexten vermehrt physikalische und chemische Konzepte abfragen, die im biologischen Zusammenhang angewendet werden müssen. Diese Transferleistung scheint eine Schwierigkeit darzustellen. In diesem Zusammenhang muss hingegen angemerkt werden, dass die Stoffwechsel-Aufgaben sowohl Diagramm- als auch Schema-Aufgaben enthielten, was in der Analyse zu den Personenfähigkeiten nicht weiter berücksichtigt wurde und einen möglichen Einfluss darstellt. Die identifizierten Unterschiede zwischen den *Repräsentationstypen* und die damit zusammenhängenden unterschiedlich vorliegenden Konventionen bekräftigen diese Annahme. Aktuelle Forschungsbefunde zeigen hingegen ein heterogenes Bild. Während die Untersuchung von Schnotz et al. (2011) zu dem Schluss kommt, dass Aufgaben mit realistischen Bildern schwieriger für Lernende zu lösen sind als Aufgaben mit Diagrammen, zeigt die Untersuchung von Brandstetter-Korinth (2017), dass Lernende realistische Abbildungen besser verstehen als abstrakte Variationen. Damit kann für die vorliegende Arbeit davon ausgegangen werden, dass sowohl Abstraktion, Konvention sowie die verwendeten Elemente und ihre Relationen zueinander eine Rolle spielen, ein heterogenes Bild diesbezüglich aber dennoch bestehen bleibt.

Die Repräsentationskompetenz und insbesondere der Umgang mit multiplen externen Repräsentationen erfordern spezifische Fähigkeiten von Lernenden, die über allgemeine sprachliche Kompetenzen hinausgehen. Vor allem der Repräsentationswechsel und die damit zusammenhängenden unterschiedlichen Translationsprozesse werden dabei von verschiedenen Faktoren beeinflusst. Wie bereits Ainsworth, Bibby, & Wood (2002) zeigen, liefert auch die vorliegende Untersuchung empirische Evidenz darüber, dass die Codierung, d.h. der Repräsentationstyp, aber auch der Abstraktionsgrad eines logischen Bildes einen Einfluss auf die Bearbeitung von Repräsentationsaufgaben haben und die Text-Bild-Integration nur dann erfolgreich bewältigt werden kann, wenn die einzelnen Darstellungsformen jeweils für sich verstanden, aufeinander bezogen und in ein mentales Modell zusammengebracht werden können (vgl. Cheng & Gilbert, 2015; Mayer, 2014; Schnotz & Bannert, 1999).

11.3 KOMPETENZ, STUDIENGANG, NOTEN UND GESCHLECHT

Die Zusammenhänge zwischen den repräsentationsspezifischen Kompetenzen MERI und MERII der Studierenden aus verschiedenen Studiengängen sowie zu den Schulnoten und dem Geschlecht wurden explorativ untersucht. Dabei zeigen die Ergebnisse, dass der Studiengang einen signifikanten Effekt auf die *MER-Integration* hat (s. Kapitel 9.4.2). Bedeutsame Unterschiede bestehen im Hinblick auf Repräsentationskompetenz hingegen nur zwischen den Studiengängen Biologie und Ingenieurwissenschaften (und ebenfalls im Kontext Stoffwechsel). Damit wird Hypothese H_{4,1} verworfen. Eine mögliche Begründung könnte in den

stark heterogen ausgerichteten Studieninhalten dieser zwei Studiengänge liegen. Bezogen auf die *Aufgabenkontexte* zeigen sich zwischen den Studiengängen Ingenieurwissenschaften und Mathematik im Kontext Genetik statistisch bedeutsame Unterschiede. Unter Berücksichtigung der beiden Teilkompetenzen *Informationsentnahmenahe Integration (MERI)* und *Konstruktionsnahe Integration (MERII)* zeigen Studierende der Mathematik die höchsten Fähigkeitswerte, wenn es um Anforderungen der MERI-Integration geht, während Studierende der Biologie die höchsten Fähigkeitswerte in den Testaufgaben zur MERII-Integration aufweisen. Fachdidaktisch relevant ist, dass diese Fähigkeitswerte mit den Aufgabenkontexten zusammenhängen. Mathematikstudierende zeigen bei der Beantwortung von Genetik-Aufgaben die höchsten sowie in den Kontexten Stoffwechsel und Ökologie die zweithöchsten Fähigkeitswerte, direkt hinter den Biologie-Studierenden. Damit liegen erste Hinweise vor, dass mathematikspezifisches Wissen zu Studienbeginn eine Rolle beim Umgang mit multiplen externen Repräsentationen spielen kann. Vor diesem Hintergrund ist auch das Ergebnis zu reflektieren, dass der Aufgabenkontext Genetik einen signifikanten Einfluss auf die Itemschwierigkeit hat. Der Einfluss ist zudem negativ (vgl. TABELLE 37, s. Kapitel 9.3.3). Genetik-Testaufgaben waren für die vorliegende Stichprobe insgesamt am leichtesten zu lösen, sind jedoch stark konventionalisiert. Dies wirft die Frage auf, inwieweit die Kenntnis und das Verstehen von Konventionen bzw. spezifischer Merkmale eines gewissen Bildtyps zu einer Reduktion der Itemschwierigkeit führen können. Vergleichende Forschungsarbeiten in diesem Bezug liegen zum aktuellen Zeitpunkt leider nur vereinzelt für die Fächer Physik oder Chemie vor und geraten erst seit kurzer Zeit in den Blick der Bildungsforschung (z.B. Pigge, Neumann, & Heinze, 2016; Walzer, Fischer, & Borowski, 2013). Die im Zusammenhang mit dem Studiengang nur schwache Varianzaufklärung ist dabei als Hinweis darauf zu interpretieren, dass der Studiengang insgesamt keinen sehr großen Einfluss auf die Personenfähigkeit hat. Damit kann davon ausgegangen werden, dass Repräsentationskompetenz am Ende der Schulzeit und als Eingangsvoraussetzung für das Studium getestet wurde. Vor diesem Hintergrund sind die eingesetzten Testaufgaben über die Fächergrenzen einsetzbar, wenn es um die Diagnostik von Kompetenzen zum Umgang mit MER geht.

Bezüglich der Zusammenhänge mit den Schulnoten liefert die vorliegende Arbeit keine statistisch bedeutsamen Ergebnisse. Zwar zeigt die Biologienote eine höchst signifikante Korrelation zu den Personenparametern, der Zusammenhang liegt jedoch im schwachen Bereich, während die Mathematiknote überhaupt nicht mit den Personenparametern korreliert (s. Kapitel 9.4.3). Damit wird die Hypothese $H_{4.2}$ bestätigt und die Hypothese $H_{4.3}$ verworfen. Aus diesem Grund scheinen die beiden Schulnoten der Fächer Biologie und Mathematik

keine entscheidende Rolle beim Lösen der vorliegenden Testaufgaben zur Repräsentationskompetenz zu spielen (vgl. auch Kapitel 10.1). Für eine solche Analyse und Bewertung sind wiederum spezifische diagnostische Kompetenzen seitens der Lehrkräfte notwendig, um repräsentationsspezifisches Wissen und die damit verbundenen Lernergebnisse zutreffend erfassen zu können und eine anschließende adäquate Notengebung sicherzustellen. Die Befunde verdeutlichen schließlich, dass die erfassten Teilkompetenzen nur in geringem Maße in die Notengebung einfließen, was wiederum aus anderen Studien der Kompetenzforschung im naturwissenschaftlichen Unterricht sowie aus PISA bekannt ist (z.B. Ziepprecht, 2016; Wellnitz, 2012; Grube, 2011; Schmiemann, 2010). Ebenfalls operationalisieren Schulnoten eher vorübergehendes und auf konkrete Prüfungen bezogenes Wissen, während die vorliegende Studie das anwendungsbezogene Wissen beim Umgang mit MER testet und diagnostiziert (vgl. z.B. Schütte et al., 2007). Zudem ist die vorliegende Stichprobe hinsichtlich der Schulnoten als insgesamt eher gut einzuschätzen, da etwa 76% der Studierenden eine Mathematiknote und etwa 84% der Studierende eine Biologienote im zweistelligen Punktebereich aufweisen, d.h. zwischen 10 und 15 Punkten. Diese Faktoren können mögliche Erklärungsansätze für die schwachen Korrelationswerte darstellen. Dennoch bleibt festzuhalten, dass Schulnoten als Indikatoren für schulische Kompetenz oftmals sachfremden Kriterien obliegen und generell als eher schlecht vergleichbar und damit kritisch zu betrachten sind (s. Kapitel 10.1). Entsprechend besitzen Schulnoten auch eine nur begrenzte Aussagekraft, wenn es um die Erfassung repräsentationsspezifischer Kompetenzen von Studierenden geht.

Zwischen weiblichen und männlichen Studierenden bestehen keine bemerkenswerten Unterschiede (s. Kapitel 9.4.3). Die weiblichen Studierenden schneiden insgesamt zwar signifikant besser ab als die männlichen Studierenden, der Unterschied ist jedoch sehr gering, wodurch Hypothese $H_{4.4}$ bestätigt wird. In diesem Zusammenhang muss hingegen erwähnt werden, dass in der vorliegenden Stichprobe fächerspezifische Unterschiede zwischen den Geschlechtern sowie Effekte in der Verteilung der Geschlechter auf die unterschiedlichen Studienganggruppen vorliegen können. Aufgrund der dadurch sehr schlechten Vergleichbarkeit wird auf eine direkte Gegenüberstellung zu anderen Studien, die sich hauptsächlich auf Schülerinnen und Schüler und nicht auf Studierende beziehen, verzichtet. Die Befunde entsprechen jedoch insgesamt dem bisherigen Forschungsstand zu Genderaspekten. So konnte die Studie von Stanat & Kunter (2001) zeigen, dass Mädchen beim Verstehen kontinuierlicher und diskontinuierlicher Texte besser abschneiden als Jungen. Die Ergebnisse decken sich ebenfalls mit Befunden der nationalen Ergänzungen im PISA-Naturwissenschaftstest, dass Mädchen im Fach Biologie einen signifikant höheren Kompetenzwert erreichen

(Baumert et al., 2001). Der Naturwissenschaftstest konnte jedoch auch zeigen, dass Jungen besser als Mädchen in der Lage sind, wenn es um die Informationsentnahme aus Diagrammen, Schemazeichnungen, Tabellen und Karten geht (Prenzel et al., 2002). Dennoch bleibt festzuhalten, dass in der vorliegenden Arbeit keine vergleichbaren kognitiven Kompetenzen erfasst und Studierende verschiedener Fachrichtungen als Stichprobe ausgewählt wurden, wodurch weitere Interpretationen in diesem Rahmen nicht sinnvoll erscheinen.

12 RELEVANZ UND AUSBLICK

Die Ergebnisse der vorliegenden Arbeit legen schulpraktische Konsequenzen nahe und liefern Ansätze für die Gestaltung von naturwissenschaftlichem Unterricht. Das Kompetenzstrukturmodell zum Umgang mit fachspezifischen Repräsentationen liefert Aussagen über die Schwierigkeit unterschiedlicher Repräsentationswechsel. Für den nachhaltigen Aufbau kommunikativer Kompetenzen ist es erforderlich, den Lernenden im naturwissenschaftlichen Unterricht grundlegende Aspekte von Repräsentationskompetenz zu vermitteln und auf die beiden unterschiedlichen Teilkompetenzen im Bereich *Informationsentnahmenaher Integration* und *Konstruktionsnaher Integration* zu fokussieren. Damit greift die Förderung von Repräsentationskompetenz an die z.B. in PISA definierten Teilkompetenzen naturwissenschaftlicher Grundbildung an (vgl. Rönnebeck et al., 2006). So ist es bei jeder MER-Integration erforderlich, die dahinterstehende naturwissenschaftliche Fragestellung zu erkennen, das naturwissenschaftliche Phänomen anhand der präsentierten Repräsentation zu erklären und durch Schlussfolgerungen aus der dargestellten Repräsentation diese als naturwissenschaftliche Evidenz zu nutzen.

Das entwickelte Kompetenzmodell kann Lehrkräfte dafür sensibilisieren, dass die Fähigkeiten zum Umgang mit multiplen externen Repräsentationen (MER) in verschiedenen Zusammenhängen kontinuierlich im Unterricht geübt werden müssen (vgl. Kozma & Russell, 2005). Damit kann bewusst am Umgang mit Visualisierungen gearbeitet und Gestaltungsprinzipien einstudiert werden (z.B. Wafi & Wirtz, 2016; Mayer, 2014; Leisen, 2011). Insbesondere die in dieser Studie festgestellten Schwierigkeiten beim Konstruieren eines vollständigen Diagramms oder einer Schemazeichnung sollten als Anlass genommen werden, verschiedene Herangehensweisen beim Umgang mit Repräsentationen explizit und reflektierend zu fördern. Auf die spezifischen Anforderungen kann sowohl im Biologieunterricht, als auch in den angrenzenden Naturwissenschaften und in der Mathematik eingegangen werden, da die Befunde zeigen, dass die Integration von multiplen externen Repräsentationen weniger vom inhaltlichen Kontext abhängig ist als von der Lese- und Übersetzungsleistung selbst. Zudem belegen Studien, dass in allen Schulbüchern der Sekundarstufe eine enorme Vielfalt unterschiedlicher Darstellungsformen existiert, die eine Förderung durch explizites Üben und Einstudieren von Techniken beim Umgang mit MER unterstreichen (z.B. Bergey et al., 2015; Griffard, 2013; Roth & Pozzer-Ardenghi, 2013). Aus didaktischer Sicht bietet es sich an, frühzeitig mit einfachen Repräsentationen zu beginnen, z.B. anhand eines Textes ein Diagramm zu konstruieren oder ein Schema zu zeichnen. Wenn Repräsentationen ohne Bezug zu den

Naturwissenschaften im Unterricht eingeführt wurden (z.B. bei Diagrammen im Mathematikunterricht), bedarf es darüber hinaus einer expliziten Klärung der zugrundeliegenden Gestaltungsmerkmale. Der Transfer der unabhängigen Variablen (x) und der abhängigen (y bzw. $f(x)$) gelingt nicht ohne weiteres auf naturwissenschaftliche Problemstellungen (von Kotzebue, Gerstl, & Nerdel 2015). Bei der Integration mehrerer Repräsentationen sollten entsprechend die einander ergänzenden Informationen herausgearbeitet und explizit aufeinander bezogen werden. In diesem Zusammenhang könnte die Entwicklung von Integrationshilfen einen Beitrag leisten, um auch den lernschwächeren Schülerinnen und Schülern die Lese- und Übersetzungsleistung Schritt für Schritt zu verdeutlichen (z.B. Corradi, Elen, & Clarebout, 2012). Vor diesem Hintergrund sollte der Förderung von Repräsentationskompetenz auch insgesamt eine höhere Aufmerksamkeit in der Schulpraxis geschenkt werden. Hierzu zählt auch, dass die unterschiedlichen Bildtypen in den Blick genommen und ihre spezifischen Herangehensweisen diskutiert werden. Dadurch kann ein aktiver Umgang mit Repräsentationswissen gesteuert werden, der zu definierten Lernergebnissen führt (vgl. Suwelack, 2010). Eine Möglichkeit könnte sein, den Umgang mit MER unter Berücksichtigung ihrer spezifischen Charakteristika und Darstellungskonventionen explizit zu thematisieren und einzuüben, da die Ergebnisse genau auf diese Anforderungen hinweisen. Gleichzeitig zeigen Studien, dass Text-Bild-Kombinationen und ihre Integration per se schwierig sind, weshalb den Lernenden vermittelt werden sollte, wie man Verknüpfungen zwischen Texten und Bildern herstellt und wie diese ineinander übersetzt werden können (z.B. Ziepprecht, 2016; Ainsworth, 2006). Werden diese Kompetenzen geübt, bergen multiple externe Repräsentationen ein großes lernförderliches Potential zum Verstehen komplexer Phänomene (Ainsworth, 1999; Krajcik & Sutherland, 2010). In diesem Zusammenhang kann ein Beitrag zum Konzeptverständnis (vgl. Ausubel, 1968) geleistet werden, indem über Repräsentationskompetenz verschiedene biologische Konzepte miteinander in Verbindung gebracht und im Sinne des Konstruktivismus eingeübt und schließlich abrufbar werden (z.B. Ummels et al., 2015; Williams & Clement, 2015). Über Kohärenzbildungsprozesse kann anschließend Strukturwissen aufgebaut werden, wodurch grundlegende Strukturen und Prinzipien erkannt und verstanden werden können (Gehl, 2013). Gerade der Integrationsprozess stellt eine besondere Herausforderung in Abituraufgaben dar, welche vor allem kohärente Lese- und Übersetzungsfähigkeiten der Schülerinnen und Schülern erwarten, die entscheidend für die anschließende Diskussion von naturwissenschaftlichen Konzepten sind.

Vor dem Hintergrund der nationalen Bildungsstandards und des Kompetenzbereichs Kommunikation sollen Lernende Texte und Bilder beschreiben und erklären, idealtypische Darstellungen sowie Schemazeichnungen, Diagramme und Symbolsprache anwenden können

(KMK, 2004). Aktuell herrscht immer noch ein Mangel an erprobten Materialien für den Biologieunterricht zum adäquaten Umgang mit Repräsentationen (Ziepprecht, 2016). Die vorliegende Arbeit bietet hierzu einen guten Ausgangspunkt für die Unterrichtsgestaltung. So liefern die konstruierten Testaufgaben weitere Gestaltungsmöglichkeiten und Fördermaßnahmen im Rahmen einer modernen Aufgabekultur, indem die Ausdifferenzierung der MER-Integration konkret für den Unterricht zur Konzeption von Lern- oder Testaufgaben herangezogen werden kann. Dadurch können Kompetenzen zu den jeweiligen Anforderungen der Repräsentationskompetenz beschrieben werden, wodurch Anwendung und Transfer in unterschiedlichen Kontexten ermöglicht wird. Gleichzeitig liefern die Ergebnisse erste Hinweise auf die Wichtigkeit der Operatoren in Aufgabenstellungen. Die Bedeutung des jeweiligen Operators und welche Anforderungen sich dahinter verbergen, muss für Lernende transparent sein und explizit eingeführt werden. Durch die Konstruktion kompetenzorientierter Lernaufgaben und eine damit verbundene Analyse vorhandener Materialien mit dem Fokus auf die genutzten Repräsentationen werden kommunikative Kompetenzen explizit gefördert und nicht nur implizit angesprochen, weil Unsicherheiten über Inhalte und Struktur der zugrundeliegenden Kompetenz bei Lehrkräften bestehen. Dadurch können die Aufgaben einerseits zur Diagnose und Rückmeldung von Schwierigkeiten im Umgang mit Repräsentationen in unterschiedlichen Themenbereichen der Biologie eingesetzt werden (z.B. Beck & Nerdel, 2016). Zudem können typische Fehler als Ausgangspunkt von Lernprozessen dienen. Dadurch werden Fördermaßnahmen durch das Reflektieren von Fehlern im Unterricht einsetzbar (z.B. Heemsoth, 2014; Schmiemann, 2014). Andererseits bieten die Aufgaben die Möglichkeit, unterschiedliche Kompetenzanforderungen bei der Integration einzuüben und zu thematisieren. In diesem Zusammenhang sollte die Vermittlung von Fachsprache und Repräsentationskompetenz in den Fokus der Lehrerausbildung und Lehrerfortbildung in Biologie rücken. Innerhalb der Aus- und Fortbildung könnte beispielsweise die vorliegende Ausdifferenzierung der MER-Integration vermittelt und anhand von Beispielen für den Transfer in den Biologieunterricht genutzt werden. Auf Seiten der angehenden Lehrkräfte sollte untersucht werden, über welche Kompetenzen diese im Bereich der MER-Integration verfügen und welches Wissen über Konventionen, Spezifika und Besonderheiten von Repräsentationen im Biologieunterricht vorhanden sind. Erst dadurch können sie die entsprechenden Hintergründe und Verständnisschwierigkeiten ihrer Schülerinnen und Schüler erkennen und lernen, diese adäquat zu diagnostizieren und mit den richtigen Hilfestellungen darauf zu reagieren (vgl. Leisen, 2011). Im Rahmen der Lehreraus- und Lehrerfortbildung können die Fachdidaktiken zur Verstärkung des Praxistransfers Aufgabenpools bereitstellen, die Inhalte zum Lehren und Lernen, ausgearbeitete Lösungsbeispiele oder Lösungshilfen und Strategien zum effektiven Ler-

nen mit Repräsentationen enthalten. Ein wichtiger Aspekt dabei sind online-Tutorials zur Unterrichts- und Aufgabenstellung, die Ergebnisse der Kompetenzmessung transparent machen und anhand der genannten Schülerfehler und den damit verbundenen richtigen Lösungsbeispielen die empirischen Befunde illustrieren. Diese Maßnahmen sollten Lehrkräfte der Naturwissenschaften in die Lage versetzen, zu beurteilen, ob sie in Klassenarbeiten und Klausuren z.B. die intendierte Fachkompetenz prüfen und nicht die Schwierigkeit durch Repräsentationen zusätzlich erhöhen. Die vorliegende Arbeit bietet hier einen entsprechenden Ausgangspunkt sowie Überblick zur Entwicklung von Lehr-, Lern- und Fördermaßnahmen, die im Unterricht implementiert werden können.

Des Weiteren bieten die Ergebnisse der Arbeit Ansatzpunkte für weitergehende Forschung. Das vorliegende, empirisch untermauerte Kompetenzstrukturmodell zum Umgang mit MER gibt detailliert Auskunft über die Dimensionen und Ausprägungen biologiespezifischer Kompetenzen in unterschiedlichen Kontexten und Bildtypen. Zudem schließt das Modell die Forschungslücke, indem speziell die Text-Bild-Integration untersucht wurde und dabei Anforderungen aus dem Biologieunterricht der Oberstufe thematisiert (vgl. Ziepprecht, 2016). In erster Linie sollte eine detaillierte Beschreibung der Teilkompetenzen vorgenommen werden. Die Analyse der Kompetenzverteilung unter Berücksichtigung des offenen Antwortformats liefert einen ersten Ausblick auf mögliche Kompetenzniveaus innerhalb der MER-Integration, die post-hoc definiert und im Rahmen weiterführender Forschung auf empirische Evidenz überprüft werden können. Durch eine solche Zuordnung von Items zu Kompetenzniveaus bzw. der domänenunabhängigen Formulierung von Anforderungsniveaus (z.B. in Anlehnung an Bybee, 2002, vgl. Terzer et al., 2013), können die kognitiven Anforderungen innerhalb der MER-Integration noch genauer spezifiziert werden. In diesem Zusammenhang stehen auch kompetenzbezogene Rückmeldungen. Anhand von Kompetenzniveaumodellen können differenzierte Bewertungen vorgenommen werden, die leistungs- und motivationsförderlicher als Rückmeldungen in Form von Schulnoten sein können (z.B. Harks et al., 2014).

Die Ergebnisse zeigen hingegen auch, dass der bisherige Forschungsstand bezüglich unterschiedlicher Bildtypen und den zugrundeliegenden Gestaltungsmerkmalen und Konventionen nach wie vor sehr heterogen ist. So kann davon ausgegangen werden, dass Abstraktion und Konvention eine Rolle spielen. Weiterführende Forschung in diesem Bereich und zu biologiespezifischen Abbildungen wie sie beispielsweise Brandstetter-Korinth (2017) liefert, wäre wünschenswert. Damit verbunden sind die Forderungen nach einer genaueren Eingrenzung und Analyse von logischen Bildern in verschiedenen biologischen Kontexten und die

Abgrenzung von abstrakten Schemazeichnungen (vgl. Brandstetter-Korinth, 2017; Ainsworth, 2006; Lohse et al., 1994). In diesem Zusammenhang sollten weiterführende Analysen zur Integration von hochkonventionalisierten Bildern durchgeführt werden. Diese könnten zusätzlich die Trennung in statistische und dynamische Bilder berücksichtigen (vgl. Brandstetter-Korinth, 2017; Wu et al., 2015; Ainsworth, Bibby, & Wood, 2002). Anhand derartiger Untersuchungen kann der Einfluss von Darstellungskonventionen auf die Kompetenzausprägungen untersucht werden. Mit den insgesamt 210 entwickelten Testaufgaben liegt zudem ein differenzierter Aufgabenpool für weiterführende Fragestellungen zur Erfassung von Kompetenzen des biologiespezifischen und naturwissenschaftlichen Repräsentationswissens vor. Die Testaufgaben im offenen Antwortformat zeigen eine größere Nähe zur Unterrichtspraxis als MC-Aufgaben, da offene Aufgaben einen Großteil der gestellten Aufgaben im Unterricht ausmachen (z.B. Jatzwauk, Rumann, & Sandmann, 2008). Aufgrund der mehrdimensionalen Skalierung lassen sich insbesondere Hinweise zur Diagnose eines relativ breiten Kompetenzspektrums sowie zur individuellen Förderung von Kompetenzen zur Übersetzungsleistung von MER ableiten. So kann der vorliegende Aufgabenpool beispielsweise für den Einsatz von Lernaufgaben genutzt werden (vgl. Jatzwauk, 2007). In diesem Zusammenhang sollte eine ausführliche qualitative Analyse der offen formulierten Antworten durchgeführt werden, um Fehlvorstellungen und typische Schwierigkeiten beim Umgang mit MER zu diagnostizieren.

Querschnittlich angelegte Untersuchungen haben den Nachteil, dass sie kaum Aussagen über Entwicklungsverläufe von Kompetenzen sowie von Einflussfaktoren und die Entstehungsursachen von Leistungsunterschieden erlauben (Grube, 2010). Im Hinblick auf die Beschreibung von repräsentationsspezifischen Kompetenzen bei der Text-Bild-Integration und damit verbundene Kompetenzzuwächse, sind Kompetenzentwicklungsmodelle nötig (Klieme & Leutner, 2006). Erst dadurch können Fördermaßnahmen über einen definierten Zeitraum ermöglicht und wirksame Maßnahmen für den Unterricht getroffen werden (Kristen et al., 2005). Aktuell fehlen empirische Ansätze zur Entwicklung und Überprüfung von Entwicklungsmodellen für den Kompetenzbereich Kommunikation, sodass hier noch ein großer Forschungsbedarf besteht. Dies betrifft ebenfalls den Einsatz von experimentellen Designs innerhalb der Interventionsforschung. Im Bereich der Biologie existieren aktuell nur vereinzelte Studien, die sich mit Unterstützungsformaten oder Instruktionsmaßnahmen beim Lernen mit Fachtexten oder z.B. Liniendiagrammen beschäftigen (z.B. Rous, 2016; Ziepprecht, 2016; Jahnke, Lumer, & Hammann, 2015). In diesem Zusammenhang sind wirkungsvolle kognitive Bearbeitungstechniken sowie Strategien für die Text-Bild-Integration notwendig, die aktuell

nur vereinzelt untersucht werden (z.B. Ziepprecht, 2016). Das Kompetenzstrukturmodell bietet Möglichkeiten der Übertragbarkeit auf weitere Fachdisziplinen im naturwissenschaftlichen Unterricht. Arbeiten, die systematisch den fachlichen Kontext über die drei Naturwissenschaften bei konstanter Art des Repräsentationswechsels betrachten, können das Kompetenzmodell fächerübergreifend einsetzbar machen. Zusätzlich könnten sich anschließende Studien auf die Untersuchung von mathematischen Voraussetzungen bei der Integration von biologiespezifischen MER fokussieren. In diesem Zusammenhang bekräftigen die Befunde die Relevanz von weiterführender Forschung, um beispielsweise die Studieneingangsvoraussetzungen im Vergleich in den Blick zu nehmen. Empfehlenswert wären Studien, die das mathematische Wissen als Eingangsvoraussetzung für ein naturwissenschaftliches Studium und für den weiteren Studienerfolg analysieren sowie Untersuchungen über die Mathematisierung als schwierigkeiterzeugender Faktor (z.B. Kimpel & Sumfleth, 2015). Dafür bieten sich Studien an, die fächerübergreifend und gemeinsam von der Naturwissenschafts- und Mathematikdidaktik erhoben werden. In diesem Zusammenhang liefern auch die ermittelten Ergebnisse erste Ansatzpunkte. Die Resultate zeigen, bezugnehmend auf die vorliegende Stichprobe und den verschiedenen Studiengängen, dass die Testaufgaben über die Fächergrenzen hinweg einsetzbar sind, wenn es um die Erfassung von Kompetenzen zum Umgang mit multiplen externen Repräsentationen geht. Ein derartiges fächerübergreifendes Arbeiten kann entsprechend zu Synergieeffekten und schließlich zu neuen wissenschaftlichen Erkenntnissen in der Hochschuldidaktik führen, wodurch Empfehlungen für die Hochschulausbildung abgeleitet werden können.

VI VERZEICHNISSE

13.1 LITERATURVERZEICHNIS

- Ackermann, T. A., Gierl, M. J., & Walker, C. M. (2003). Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51.
- Adams, R. J. (2005). Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, 31, 162-172.
- Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, 21(1), 1-23.
- Adams, R. J., & Wu, M. (2002). *PISA 2000. Technical Report*. Paris: OECD.
- Adams, R. J., & Wu, M. (2010). *Unidimensional Latent Regression*. Zugriff am 07.07.2017. Verfügbar unter <https://www.acer.org/files/Conquest-Tutorial-8-MultidimensionalLatent-Regression.pdf>.
- Ainsworth, S. E. (1999). The functions of multiple representations. *Computers and Education*, 33(2-3), 131-152.
- Ainsworth, S. E. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3), 183-198.
- Ainsworth, S. E., & Th Loizou, A. (2003). The effects of self-explanation when learning with text or diagrams. *Cognitive Science*, 27, 669-681.
- Ainsworth, S.E., Bibby, P. A., & Wood, D. J. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences*, 11, 25-62.
- Ainsworth, S. E., Prain, V., & Tytler, R. (2011). Drawing to Learn in Science. *Science*, 333(6046), 1096-1097.
- Altman, D. G. (1991). *Practical Statistics for medical research*. London: Chapman & Hall.
- Amelang, M., & Zielinski, W. (2002). *Psychologische Diagnostik und Intervention*. Berlin [u.a.]: Springer.
- Anderson, T. R., Schönborn, K. J., du Plessis, L., Gupthar, A. S., & Hull, T. L. (2013). Identifying and Developing Students' Ability to Reason with Concepts and Representations in Biology. In D. F. Treagust & C.-Y. Tsui (Hrsg.), *Multiple Representations in Biological Education* (S. 19-38). Dordrecht [u.a.]: Springer.

- Anzai, Y. (1991). Learning and use of representations for physics expertise. In K. A. Ericsson & J. Smith (Hrsg.), *Toward a general theory of expertise – Prospects and limits* (S. 64-92). New York: Cambridge University Press.
- Arnheim, R. (1972). *Anschauliches Denken*. Köln: DuMont Schuberg.
- Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69-137). Opladen: Leske + Budrich.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In K. W. Spence & J. T. Spence (Hrsg.), *The Psychology of Learning and Motivation* (S. 89-195). New York: Academic Press.
- Atkinson, R. C., & Shiffrin, R. M. (1971). The Control Processes of Short-term Memory, Technical Report 173, April 19. *Psychology Series, Stanford University*. Stanford: Institute for Mathematical Studies in the Social Sciences.
- Ausubel, D. P. (1968). *Educational Psychology: A Cognitive View*. New York: Holt, Rinehart and Winston, Inc.
- Ayres, P., & Sweller, J. (2005). The split-attention-principle in multimedia learning. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (S. 135-146). New York: Cambridge University Press.
- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*. Berlin: Springer.
- Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.
- Baddeley, A. (1992). Working memory. *Science*, 255, 556-559.
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neurosciences*, 4(10), 829-839.
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2001). Toward a Model of Learning Data Representations. In J. D. Moore & K. Stenning (Hrsg.), *Proceedings of the 23rd Conference of the Cognitive Science Society* (S. 45-50). NJ: Lawrence Erlbaum Associates.
- Ballstaed, S.-P., Mandl, H., Schnotz, W., & Tergan, S.-O. (1981). *Texte verstehen. Texte gestalten*. München: Urban & Schwarzenburg.
- Baumert, J., Bos, W., & Lehmann, R. (2000). TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Mathematische und naturwissenschaftliche Grundbildung am Ende der gymnasialen Oberstufe, Bd. 2. Opladen: Leske+Budrich.

- Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2001). 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske+Budrich.
- Baumert, J., Stanat, P., & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15-68). Opladen: Leske+Budrich.
- Bayrhuber, H., Bögeholz, S., Elster, D. et al. (2007). Biologie im Kontext. Erste Forschungsergebnisse. *Der mathematische und naturwissenschaftliche Unterricht*, 60(5), 304-313.
- Beck, C., & Nerdel, C. (2016). Bildverständnis im Biologieunterricht diagnostizieren und fördern. *Unterricht Biologie*, 417, 38-43.
- Bennett, J., Hogarth, S., & Lubben, F. (2003). *A systematic review of the effects of context-based and Science-Technology-Society (STS) approaches in the teaching of secondary science*. Social Science Research Unit, Institute of Education, University of London.
- Berck, K.-H., & Graf, D. (1992). Begriffsauswahl und Begriffsvermittlung – Überblick über den Forschungsstand für den Biologieunterricht. In H. Entrich & L. Staeck (Hrsg.), *Sprache und Verstehen im Biologieunterricht* (S. 76-90). Bad Zwischenahn: Leuchtturm-Verlag.
- Bernholt, S. (2010). *Kompetenzmodellierung in der Chemie – Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität* (Studien zum Physik- und Chemielernen, Bd. 98). Berlin: Logos Verlag.
- Bernholt, S., Parchmann, I., & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. *Zeitschrift für Didaktik der Naturwissenschaft*, 15, 219-245.
- Bergey, B. W., Cromley, J. G., & Newcombe, N. S. (2015). Teaching High School Biology Students to Coordinate Text and Diagrams: Relations with Transfer, Effort, and Spatial Skill. *International Journal of Science Education*, 37(15), 2476-2502.
- Biological Sciences Curriculum Study (2006) *BSCS Biology: A human approach (teacher guide)* (3rd. Ed.). Dordrecht, The Netherlands: Springer.
- Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 792-810.
- Blömeke, S. (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und –referendare. Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Münster [u.a.]: Waxmann.
- Bodemer, D. (2004). *Unterstützung multimedialen Lernens: Instruktionsdesign zur Förderung der Kohärenzbildung beim Wissenserwerb mit Texten und dynamisch-interaktiven Bildern*. Dissertation, Albert-Ludwigs-Universität Freiburg, Freiburg.

- Bodemer, D., & Faust, U. (2006). External and mental referencing of multiple representations. *Computers in Human Behavior*, 22(1), 27-42.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences* (2. Auflage). Mahwah [u.a.]: Lawrence Erlbaum Associates, Publishers.
- Boone, W. J., & Scantlebury, K. (2006). The Role of Rasch Analysis When Conducting Science Education Research Utilizing Multiple-Choice-Tests. *Science Education*, 90, 253-269.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg [u.a.]: Springer.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg [u.a.]: Springer.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (Psychologie, 7. Auflage). Heidelberg [u.a.]: Springer.
- Brandstetter-Korinth, M. (2017). *Abbildungen im Biologieunterricht* (Biologie lernen und lehren, Bd. 19). Berlin: Logos Verlag.
- Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling*, 54(1), 36-53.
- Brosius, F. (2013). *SPSS 21*. Heidelberg: mitp.
- Brünken, R., Seufert, T., & Zander, S. (2005). Förderung der Kohärenzbildung beim Lernen mit multiplen Repräsentationen. *Zeitschrift für Pädagogische Psychologie*, 19(1/2), 61-75.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (Psychologie, 2., Auflage). München: Pearson Studium.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (Psychologie, 3., aktual. und erw. Auflage). München: Pearson Studium.
- Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler* (Ps Psychologie). München: Pearson Studium.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodal Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33, 261-304.
- Bybee, R. W., & McCrae, B. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education*, 33, 7-26.
- Bybee, R. W. (2002). Scientific Literacy – Mythos oder Realität? In W. Gräber, P. Nentwig, T. Koballa & R. Evans (Hrsg.), *Scientific Literacy. Ein Beitrag der Naturwissenschaften zur Allgemeinen Bildung* (S. 21-43). Opladen: Leske & Budrich.

- Bybee, R. W. (1997). *Achieving scientific literacy: From puposes to practices*. Portsmouth, NH: Heinemann.
- Carstensen, C. H., Knoll, S., Frey, A., & Walter, O. (2007). Technische Grundlagen des dritten internationalen Vergleichs. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 367-390). Münster: Waxmann.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293-332.
- Cheng, M. M., & Gilbert, J. K. (2015). Students' Visualization of Diagrams Representing the Human Circulatory System: The use of Spatial Isomorphism and Representational Conventions. *International Journal of Science Education*, 37(1), 136-161.
- Cheng, M., & Gilbert, J. K. (2009). Towards a Better Utilization of Diagrams in Research into the Use of Representative Levels in Chemical Education. In J. K. Gilbert & D. Treagust (Hrsg.), *Multiple Representations in Chemical Education: Models and Modeling in Science Education* (S. 55-73). Dordrecht: Springer.
- Chung, Y., Yoo, J., Kim, S.-W., Lee, H., & Zeidler, D. L. (2016). Enhancing Students' Communication Skills in the Science Classroom through Socioscientific Issues. *International Journal of Science and Mathematics Education*, 14(1), 1-27.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Auflage). Hillsdale [u.a.]: Lawrence Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Connell, M. W., Sheridan, K., & Gardner, H. (2003). On abilities and domains. In R. J. Sternberg & E. L. Grigorenko (Hrsg.), *The psychology of abilities, competencies, and expertise* (S. 126-155). Cambridge: Cambridge University Press.
- Cook, M. P. (2006). Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education*, 90(6), 1073-1091.
- Cook, M., Wiebe, E. N., & Carter, G. (2008). The influence of prior knowledge on viewing graphs with macroscopic and molecular representations. *Science Education*, 92, 848-867.
- Corradi, D., Elen, J., & Clarebout, G. (2012). Understanding and enhancing the use of multiple external representations in chemistry education. *Journal of Science Education and Technology*, 21(6), 780-795.
- Corradi, D., Elen, J., Schraepen, B., & Clarebout, G. (2014). Understanding possibilities and limitations of abstract chemical representations for achieving conceptual understanding. *International Journal of Science Education*, 36(5), 715-734.

- Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and Instruction*, 9(4), 343-363.
- Cromley, J. G., Bergey, B. W., Fitzhugh, S., Newcombe, N., Wills, T. W., Shipley, T. F., & Tanaka, J. C. (2013). Effects of three diagram instruction methods on transfer of diagram comprehension skills: The critical role of inference while learning. *Learning and Instruction*, 26, 45-58.
- Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology*, 35(1), 59-74.
- Crowell, A., & Schunn, C. (2015). Unpacking the Relationship Between Science Education and Applied Scientific Literacy. *Research in Science Education*, 45(2), 119-140.
- DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, 37(6), 582-601.
- Delgado, C., & Lucero, M. M. (2015). Scale construction for graphing: An investigation of students' resources. *Journal of Research in Science Teaching*, 52(5), 633-658.
- Deregowski, J. B. (1989). Real space and represented space: Cross-cultural perspectives. *Behavioral and Brain Sciences*, 12, 15-119.
- Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
- Downing, S. M, & Haladyna, T. M. (1997). *Handbook of Test Development*. Mahwah [u.a.]: Lawrence Erlbaum Associates, Publishers.
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Auflage). Berlin [u.a.]: Springer.
- Dreyfus, T., & Eisenberg, t. (1990). On difficulties with diagrams: Theoretical issues. In G. Booker, P. Coob & T. N. De Mendicuti (Hrsg.), *Proceedings of the 14th Annual Conference of the International Group for the Psychology of Mathematics Education* (Vol 1, S. 27-36). Oaxtepec: PME.
- Duit, R., Häußler, P., & Prenzel, M. (2001). Schulleistungen im Bereich der naturwissenschaftlichen Bildung. In F. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 169-186). Weinheim: Beltz.
- Dytham, C. (2005). *Chosing and Using Statistics: A Biologist's Guide (3rd Edition)*. Blackwell Science.
- Eggert, S. & Bögeholz, S. (2010). Students' Use of Decision Making Strategies with regard to Socioscientific Issues – An Application of the Rasch Partical Credit Model. *Science Education*, 94, 230-258.

- Eilam, B. (2012). *Teaching, Learning, and Visual Literacy. The Dual Role of Visual Representation*. Cambridge: Cambridge University Press.
- Eilam, B., & Poyas, Y. (2010). External Visual Representations in Science Learning: The case of relations among system components. *International Journal of Science Education*, 32(7), 2335-2366.
- Einhaus, E. A. (2007). *Schülerkompetenzen im Bereich Wärmelehre. Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen* (Studien zum Physik- und Chemielernen, Bd. 63). Berlin: Logos Verlag.
- Eitel, A., & Scheiter, K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational Psychology Review*, 27(1), 153-180.
- Embretson, S. E., & Reise, S. (2000). *Item Response Theory For Psychologists* (Multivariate Applications). Mahwah [u.a.]: Lawrence Erlbaum Associates, Publishers.
- Enzingmüller, C., Nerdel, C., & Prechtel, H. (2012). Lehrerüberzeugungen zur Fachsprache im Biologieunterricht. In M. Gläser-Zikuda, T. Seidel, C. Rohlf, A. Gröschner & S. Ziegelbauer (Hrsg.), *Mixed Methods in der empirischen Bildungsforschung* (S. 183-194). Münster [u.a.]: Waxmann.
- Ferber, N. (2014). *Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I* (Studien zum Physik- und Chemielernen, Bd. 169). Berlin: Logos Verlag.
- Field, A. P. (2009). *Discovering Statistics Using SPSS. (and sex and drugs and rock 'n' roll)* (Introduction statistical methods, 3. Auflage). Los Angeles [u.a.]: SAGE Publications.
- Finkelstein, N. (2005). Learning physics in context: A study of student learning about electricity and magnetism. *International Journal of Science Education*, 27, 1187-1209.
- Fischer, H. E., & Draxler, D. (2001). Aufgaben und naturwissenschaftlicher Unterricht. *Der mathematische und naturwissenschaftliche Unterricht*, 54, 388-393.
- Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.
- Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunkt-programms. *Zeitschrift für Erziehungswissenschaft*, 16(1), 5-22.
- Fleischer, J., Wirth, J., & Leutner, D. (2014). Effekte der kontextuellen Einkleidung von Testaufgaben auf die Schülerleistungen im analytischen Problemlösen und in der Mathematik. *Zeitschrift für Pädagogische Psychologie*, 28, 207-227.
- Florian, C. (2012). Abituraufgaben im Fach Biologie. Schwierigkeitsbestimmende Merkmale schriftlicher Prüfungsaufgaben (Biologie lernen und lehren, Bd. 4). Berlin: Logos Verlag.

- Florian, C., Sandmann, A., & Schmiemann, P. (2014). Modellierung kognitiver Anforderungen schriftlicher Abituraufgaben im Fach Biologie. *Zeitschrift für Didaktik der Naturwissenschaften*, 20(1), 175-189.
- Florian, C., Schmiemann, P., & Sandmann, A. (2015). Aufgaben im Zentralabitur Biologie – eine kategoriengestützte Analyse charakteristischer Aufgabenmerkmale schriftlicher Abituraufgaben. *Zeitschrift für Didaktik der Naturwissenschaften*, 21(1), 69-86.
- Frey, A., & Annageldyev, M. (2015). Youden. A program for the construction of booklet designs (Version 1.0) [Computer Software]. Jena: Friedrich Schiller University, Germany.
- Frey, A., & Hartig, J. (2009). Editorial: Assessment of competencies. *Studies in Educational Evaluation*, 35, 55-56.
- Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice* 28(3), 39-53.
- Gehl, D. (2013). *Vom Betrachten zum Verstehen. Eine Diagnose von Rezeptionsprozessen und Wissensveränderungen bei multimodalen Printclustern*. Wiesbaden: Springer VS.
- Gilbert, J. K. (2007). Visualization: A metacognitive skill in science and science education. In J. K. Gilbert (Hrsg.), *Visualization in Science Education* (1. Auflage, S. 9-27). Dordrecht, The Netherlands: Springer.
- Gilbert, J. K. (2006). On the Nature of "Context" in Chemical Education. *International Journal of Science Education*, 28(9), 957-976.
- Gilbert, J. K., & Treagust, D. F. (2009). Introduction: Macro, Submicro and Symbolic Representations and the Relationship Between Them: Key Models in Chemical Education. In J. Gilbert & D. F. Treagust (Hrsg.), *Multiple Representations in Chemical Education. Models and Modeling in Science Education* (S. 1-8). Dordrecht: Springer.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237-288.
- Goldmann, S. R., & Bisanz, G. L. (2002). Toward a Functional Analysis of Scientific Genres: Implications for Understanding and Learning Processes. In J. Otero, J. A. León & A. C. Graesser (Hrsg.), *The Psychology of Science Text Comprehension* (S. 19-50). Mahwah: Lawrence Erlbaum.
- Gollwitzer, M. (2008). Latent-Class-Analysis. In H. Moosbrugger & A. Kelava (Hrsgt), *Testtheorie und Fragebogenkonstruktion* (S. 279-306). Heidelberg: Springer.
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30(5), 349-411.

- Graf, D. (2001). Welche Aufgabentypen gibt es? *Der mathematische und naturwissenschaftliche Unterricht*, 54, 422-425.
- Green, S. B. (1991). How many Subjects does it take to do a Regression Analysis? *Multivariate Behavioral Research*, 26(3), 499-510.
- Greve, W., & Wentura, D. (1997). *Wissenschaftliche Beobachtung: Eine Einführung*. Weinheim: PVU/ Beltz.
- Griffard, P. B. (2013). Deconstructing and Decoding Complex Process Diagrams in University Biology. In D. F. Treagust & C.-Y. Tsui (Hrsg.), *Multiple Representations in Biological Education* (S. 165-184). Dordrecht [u.a.]: Springer.
- Grouven, U., Bender, R., Ziegler, A., & Lange, S. (2007). Der Kappa-Koeffizient. *Deutsche Medizinische Wochenschrift*, 132(1), 65-68.
- Grube, C. (2011). *Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung. Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I*. Dissertation, Universität Kassel, Kassel.
- Habermas, J. (1971). Vorbereitende Bemerkungen zu einer Theorie der kommunikativen Kompetenz. In J. Habermas & N. Luhmann (Hrsg.), *Theorie der Gesellschaft oder Sozialtechnologie – Was leistet die Systemforschung?* (S. 101-141). Frankfurt a. M.: Suhrkamp.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (3. Auflage). Mahwah [u.a.]: Lawrence Erlbaum Associates.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on Educational Measurement: issues and practice. *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*, 13(3), 38-47.
- Hammann, M. (2004). Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung – dargestellt anhand von Kompetenzen beim Experimentieren. *Mathematischer und Naturwissenschaftlicher Unterricht*, 57(4), 196-203.
- Hammann, M., & Jördens, J. (2014). Offene Aufgaben codieren. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 169-178). Berlin [u.a.]: Springer.
- Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology*, 34(3), 269-290.
- Harms, U., Mayer, J., Hammann, M., Bayrhuber, H., & Kattmann, U. (2004). Kerncurriculum und Standards für den Biologieunterricht in der gymnasialen Oberstufe. In H. E. Tenorth (Hrsg.), *Kerncurriculum für die Oberstufe II* (S. 22-84). Weinheim: Beltz.

- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung; DESI-Studie (Deutsch-Englisch-Schülerleistungen-International)* (Beltz Pädagogik, S. 83-99). Weinheim [u.a.]: Beltz.
- Hartig, J. (2008). Psychometric Models for the Assessment of Competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 69-90). Göttingen: Hogrefe & Huber Publishers.
- Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, 63, 43-49.
- Hartig, J., & Frey, A. (2013). Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl“ für die Modellierung von Kompetenzen? *Zeitschrift für Erziehungswissenschaft*, 16 (Sonderheft), 47-51.
- Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143-171). Heidelberg: Springer.
- 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* (56. Beiheft der Zeitschrift für Pädagogik, S. 189-198). Weinheim & Basel: Beltz.
- Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127-143). Berlin [u.a.]: Springer.
- Hartig, J., & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values“ in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merckens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 27-44). Wiesbaden [u.a.]: VS Verlag für Sozialwissenschaften.
- Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 17-36). Bonn [u.a.]: Bundesministerium für Bildung und Forschung.
- Hartig, J., Jude, N., & Wagner, W. (2008). Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen. In E. Klieme (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (Beltz Pädagogik, S. 33-54). Weinheim [u.a.]: Beltz.
- Hartig, J., & Rauch, D. (2010). Multiple-Choice versus Open-Ended Response Format of Reading Test Items: A Two-Dimensional IRT-Analysis. *Psychological Test and Assessment Modeling*, 52(4), 354-379.
- Hartig, H., Bernholt, S., & Retelsdorf, J. (2015). Fachsprache oder Sprache im Fach – Stand der Forschung und Forschungsperspektiven zur Rolle der Sprachen im naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 21, 55-67.

- Härtig, H., Heitmann, P., & Retelsdorf, J. (2015). Analyse der Aufgaben zur Evaluation der Bildungsstandards in Physik – Differenzierung von schriftsprachlichen Fähigkeiten und Fachlichkeit. *Zeitschrift für Erziehungswissenschaft*, 18, 763-779.
- Haugwitz, M. (2009). *Kontextorientiertes Lernen und Concept Mapping im Fach Biologie. Eine experimentelle Untersuchung zum Einfluss auf Interesse und Leistung unter Berücksichtigung von Moderationseffekten individueller Voraussetzungen beim kooperativen Lernen*. Dissertation, Universität Duisburg-Essen. Essen.
- Heemsoth, T. (2014). Fehler als Lerngelegenheiten im Mathematikunterricht. *IPN-Blätter-Informationen aus dem Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik*, 3/31, 7.
- Hegarty, M. (2011). The Cognitive Science of Visual-Spatial Displays: Implications for Design. *Topics in Cognitive Science*, 3(3), 446-474.
- Hegarty, M., Carpenter, P. A., & Just, M. A. (1991). Diagrams in the comprehension of scientific texts. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Hrsg.), *Handbook of Reading Research* (Vol. 2, S. 641-668). New York: Longman.
- Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of memory and language*, 32(6), 717-742.
- Hettmannsperger, R. (2015). *Lernen mit multiplen Repräsentationen aus Experimenten. Ein Beitrag zum Verstehen physikalischer Konzepte*. Wiesbaden: Springer VS.
- Hochpöchler, U., Schnotz, W., Rasch, T., Ullrich, M., Horz, H., McElvany, N., ... Baumert, J. (2013). Dynamics of mental model construction from text and graphics. *European Journal of Psychology of Education*, 28(4), 1105-1126.
- Hohensinn, C., & Kubinger, K. D. (2011). Applying Item Response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*, 71, 732-746.
- Hubber, P., Tytler, R., & Haslam, F. (2010). Teaching and Learning about Force with a Representational Focus: Pedagogy and Teacher Change. *Research in Science Education*, 40, 5-28.
- Jahnke, L., Lumer, J., & Hammann, M. (2015). Förderung des Beschreibens und Erklärens von Liniendiagrammen im Biologieunterricht. *Paper presented at 17. Frühjahrsschule der Fachsektion Didaktik der Biologie*. München.
- Jankisz, E., & Moosbrugger, H. (2008). Planung und Entwicklung von psychologischen Tests und Fragebögen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 28-71). Heidelberg: Springer.
- Janssen, J., & Laatz, W. (2005). *Statistische Datenanalyse mit SPSS. Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests* (5. Auflage). Berlin [u.a.]: Springer.

- Jatzwauk, P. (2007). *Aufgaben im Biologieunterricht. Eine Analyse der Merkmale und des didaktisch-methodischen Einsatzes von Aufgaben im Biologieunterricht*. Berlin: Logos-Verlag.
- Jatzwauk, P., Rumann, S., & Sandmann, A. (2008). Der Einfluss des Aufgabeneinsatzes im Biologieunterricht auf die Lernleistung der Schüler – Ergebnisse einer Videostudie. *Zeitschrift für Didaktik der Naturwissenschaft*, 14, 263-283.
- Johnstone, A. H. (1982). Macro and micro chemistry. *School Science Review*, 19(3), 71-73.
- Jörissen, S., & Schmidt-Thieme, B. (2015). Darstellen und Kommunizieren. In R. Bruder et al. (Hrsg.), *Handbuch Mathematikdidaktik* (S.385-408). Berlin [u.a.]: Springer.
- Kalverkämper, H. (1998). Fachsprache und Fachsprachenforschung. In L. Hoffmann, H. Kalverkämper & H. E. Wiegand (Hrsg.), *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* (S. 48-59). Berlin: DeGruyter.
- Kampa, N. N. (2012). *Aspekte der Validierung eines Tests zur Kompetenz in Biologie. Eine Studie zur Kompetenz in Biologie und ihren Teildimensionen Konzept- und Prozesswissen*. Dissertation, Humboldt-Universität zu Berlin. Berlin.
- Kaput, J. J. (1989). Linking representations in the symbol systems of algebra. In S. Wagner & C. Kieran (Hrsg.), *Research issues in the learning and teaching of algebra* (S. 167-194). Hillsdale, NJ: Erlbaum.
- Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben* (Studien zum Physik- und Chemilernen, Bd. 79). Berlin: Logos Verlag.
- Kauertz, A. (2014). Entwicklung eines Rasch-skalierten Leistungstest. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 341-353). Berlin [u.a.]: Springer.
- Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, 16, 135-153.
- Kelava, A., & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 75-102). Heidelberg: Springer.
- Kimpel, L., & Sumfleth, E. (2015). Chemieaufgaben: Mathematisierung als schwierigkeitserzeugender Faktor. In C. Maurer (Hrsg.), *Authentizität und Lernen – das Fach in der Fachdidaktik: Gesellschaft für Didaktik der Chemie und Physik (Jahrestagung in Berlin 2015, 14.-17.09.2015)*. S. 521-523.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

- Kintsch, W., & Dijk, T. A. van (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.
- Kirsh, D. (2009). Problem solving and situated cognition. In P. Robbins & M. Aydede (Hrsg.), *The cambridge handbook of situated cognition* (S. 264-306). Cambridge: Cambridge University Press.
- Klauer, K. (1986). Kriteriumsorientiertes Testen: Der Schluss auf den Itempool. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 8, 141-147.
- Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Bd. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 57-128). Opladen: Leske + Budrich.
- Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (2010). PISA 2009. Bilanz nach einem Jahrzehnt. Münster [u.a.]: Waxmann.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Expertise* (Bildungsforschung, Bd. 1). Bonn.
- Klieme, E., Baumert, J., Köller, O., & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung. Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos, R. Lehmann (Hrsg.), *TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit, Bd. 1*. Opladen: Leske+Budrich.
- Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 3-22). Cambridge, MA: Hogrefe & Huber.
- Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, 52(6), 876-903.
- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222-237.
- Klieme, E., & Steinert, B. (2004). Einführung der KMK-Bildungsstandards. *Der mathematische und naturwissenschaftliche Unterricht*, 57(3), 132-137.
- Klöpffel, K., Schwanewedel, J., & Mayer, J. (2013). Biologische Informationen aus unterschiedlichen Repräsentationen erschließen. Entwicklung und Überprüfung eines Testinstruments. In D. Krüger, P. Schmiemann, A. Möller, A. Dittmer & J. Zabel (Hrsg.), *Erkenntnisweg Biologiedidaktik 12* (S. 27-42). Berlin: FU Berlin.

- KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesregierung Deutschland. (2005a). *Beschlüsse der Kultusministerkonferenz – Bildungsstandards im Fach Biologie für den mittleren Bildungsabschluss (Beschluss vom 16. Dezember 2004)*. München: Wolters Kluwer.
- KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesregierung Deutschland. (2005b). *Beschlüsse der Kultusministerkonferenz – Bildungsstandards im Fach Chemie für den mittleren Bildungsabschluss (Beschluss vom 16. Dezember 2004)*. München: Wolters Kluwer.
- KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesregierung Deutschland. (2005c). *Beschlüsse der Kultusministerkonferenz – Bildungsstandards im Fach Physik für den mittleren Bildungsabschluss (Beschluss vom 16. Dezember 2004)*. München: Wolters Kluwer.
- KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesregierung Deutschland. (2004). *Einheitliche Prüfungsanforderungen in der Abiturprüfung. Biologie (Beschluss vom 01.12.1979 i.d.F. vom 24.10.2008)*. Luchterhand: München.
- Knigge, J. (2010). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“*. Dissertation, Universität Bremen, Bremen.
- Köller, O., Fischer, H. E., Mayer, J., Sumfleth, E., Hartmann, S., Katzenbach, M., Kauertz, A., Notarp, H., Ropohl, M., Walpuski, M., Wellnitz, N., & Zilker, I. (2008a). Evaluation der Standards in den Fächern Biologie, Chemie und Physik für die Sekundarstufe I (ESNaS). Band 1: Projektbeschreibung.
- Kosslyn, S. M. (1989). Understanding Charts and Graphs. *Applied Cognitive Psychology*, 3(3), 185-226.
- Kotzebue, L. von (2014). *Diagrammkompetenz als biologiedidaktische Aufgabe für die Lehrerbildung: Konzeption, Entwicklung und empirische Validierung eines Strukturmodells zum diagrammspezifischen Professionswissen im biologischen Kontext*. Dissertation, Technische Universität München, München.
- Kotzebue, L. von, Gerstl, M., & Nerdel, C. (2015). Common Mistakes in the Construction of Diagrams in Biological Contexts. *Research of Science Education*, 45(2), 193-213.
- Kotzebue, L. von, & Nerdel, C. (2012). Professionswissen von Biologielehrkräften zum Umgang mit Diagrammen. *Zeitschrift für Didaktik der Naturwissenschaften*, 18, 181-200.
- Kotzebue, L. von, & Nerdel, C. (2015). Modellierung und Analyse des Professionswissen zur Diagrammkompetenz bei angehenden Biologielehrkräften. *Zeitschrift für Erziehungswissenschaft*, 1-26.
- Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction*, 13(2), 205-226.

- Kozma, R., Chin, E., Russell, J., & Marx, N. (2000). The roles of representations and tools in the chemistry laboratory and their implications for chemistry learning. *Journal of the Learning Sciences*, 9(2), 105-143.
- Kozma, R., & Russell, J. (2005). Students becoming chemists: Developing representational competence. In J. K. Gilbert (Ed.), *Visualizations in science education* (pp. 121-146). Dordrecht, The Netherlands: Springer.
- Kozma, R., & Russell, J. (1997). Multimedia and Understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, 34(9), 949-968.
- Krajcik, J. S., & Sutherland, L. M. (2010). Supporting students in developing literacy in science. *Science*, 328(5977), 456-459.
- Kragten, M., Admiraal, W., & Rijlaarsdam, G. (2013). Diagrammatic literacy in secondary science education. *Research in Science Education*, 43(5), 1785-1800.
- Krapp, A., & Prenzel, M. (1992). *Interesse, Lernen, Leistung. Neuere Ansätze einer pädagogisch-psychologischen Interessenforschung*. Münster: Aschendorff.
- Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods and findings. *International Journal of Science Education*, 33(1), 27-50.
- Kristen, C., Römmer, A., Müller, W., & Kalter, F. (2005). *Längsschnittstudien für die Bildungsberichterstattung: Beispiele aus Europa und Nordamerika*. Berlin: Bundesministerium für Bildung und Forschung.
- Krüger, D., & Riemeier, T. (2014). Die qualitative Inhaltsanalyse – eine Methode zur Auswertung von Interviews. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 133-146). Berlin [u.a.]: Springer.
- Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, 18, 111-115.
- Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science*, 49, 361-374.
- Kuha, J. (2004). AIC and BIC. Comparisons of Assumptions and Performance. *Sociological Methodes & Research*, 33(2), 188-229.
- Kulgemeyer, C. (2010). *Physikalische Kommunikationskompetenz. Modellierung und Diagnostik* (Studien zum Physik- und Chemielernen, Bd. 108). Berlin: Logos Verlag.

- Kulgemeyer, C., & Schecker, H. (2009). Kommunikationskompetenz in der Physik: Zur Entwicklung eines domänenspezifischen Kommunikationsbegriffs. *Zeitschrift für Didaktik der Naturwissenschaften*, 15, 131-153.
- Kulgemeyer, C., & Schecker, H. (2007). PISA 2000 bis 2006 – Ein Vergleich anhand eines Strukturmodells für naturwissenschaftliche Aufgaben. *Zeitschrift für Didaktik der Naturwissenschaften*, 13, 199-220.
- Kürschner, C., & Schnotz, W. (2007). Konstruktion mentaler Repräsentationen bei der Verarbeitung von Text und Bildern. *Unterrichtswissenschaften*, 35(1), 48-67.
- Labudde, P. (2000). *Konstruktivismus im Physikunterricht der Sekundarstufe II*. Bern: Paul Haupt Verlag.
- Lachmayer, S. (2008). *Entwicklung und Überprüfung eines Strukturmodells der Diagrammkompetenz für den Biologieunterricht*. Elektronische Dissertation, Christian-Albrechts-Universität, Kiel, Germany. Zugriff am 21.06.2017. Verfügbar unter <http://dnb.info/1019667389/34>.
- Lachmayer, S., Nerdel, C., & Prectl, H. (2007). Modellierung kognitiver Fähigkeiten beim Umgang mit Diagrammen im naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 13, 145-160.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research*, 60(1), 1-64.
- Leisen, J. (2005). Wechsel der Darstellungsform. Eine wichtige Strategie im kommunikativen Physikunterricht. *Unterricht Physik*, 87, 10-11.
- Leisen, J. (2006). Aufgabenkultur im mathematisch-naturwissenschaftlichen Unterricht. *Der mathematische und naturwissenschaftliche Unterricht*, 59(5), 260-266.
- Leisen, J. (2011). Praktische Ansätze schulischer Sprachförderung – Der sprachensible Fachunterricht. Zugriff am 04.07.2017. Verfügbar unter https://www.hss.de/fileadmin/media/downloads/Berichte/111027_RM_Leisen.pdf.
- Leisen, J. (2015). Fachlernen und Sprachlernen! *Verband zur Förderung des MINT-Unterrichts*, 68(3), 132-137.
- Lem, S., Kempen, G., Ceulemans, E., Onghena, P., Verschaffel, L., & van Dooren, W. (2015). Combining Multiple External Representations and Refutational Text: An Intervention on Learning to Interpret Box Plots. *International Journal of Science and Mathematics Education*, 1-18.
- Lemke, J. L. (1990). *Talking Science: Language, Learning, and Values*. Norwood, NJ: Ablex Publishing Corporation.

- Lemke, J. L. (1998b). *Teaching all the languages of science: Words, symbols, images, and actions*. Paper presented at the International Conference on Ideas for a Scientific Culture, Barcelona, Spain. Zugriff am 21.06.2017. Verfügbar unter <http://academic.brooklyn.cuny.edu/education/jlemke/papers/barcelon.htm>.
- Lemke, J. L. (2004). The literacies of science. In E. W. Saul (Hrsg.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (S. 33-47). Arlington, VA: International Reading Association.
- Leucht, M., Retelsdorf, J., Möller, J., & Köller, O. (2010). Zur Dimensionalität rezeptiver englischsprachiger Kompetenzen. *Zeitschrift für Pädagogische Psychologie*, 24, 123-138.
- Leutner, D., Hartig, J., & Jude, N. (2008). Measuring Competencies: Introduction to Concepts and Questions of Assessment in Education. In E. Klieme, D. Leutner & J. Hartig (Hrsg.), *Assessment of competencies in educational contexts* (S. 177-192). Toronto: Hogrefe & Huber Publishers.
- Levie, H., & Lentz, R. (1982). Effects of text illustrations: A review of research. *Educational Communication and Technology Journal*, 30(4), 195-232.
- Levin, J. R., Anglin, G. J., & Carney, R. N. (1987). On empirically validating functions of pictures in prose. In D. M. Willows & H. A. Houghton (Hrsg.), *The psychology of illustration*, 51-85.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Auflage). Weinheim: Beltz.
- Linacre, J. M. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions*, 7(4), 328.
- Lind, G., & Friege, G. (2003). Wissen und Problemlösen. Eine Untersuchung zur Frage des "trägen Wissens". *Empirische Pädagogik*, 17(1), 57-86.
- Lind, G., Friege, G., & Sandmann, A. (2005). Selbsterklären und Vorwissen. *Empirische Pädagogik*, 19(1), 1-27.
- Lohse, G. L., Biolski, K., Walker, N., & Schmidt, J. D. (1994). A Classification of Visual Representations. *Communications of the ACM*, 37(12), 36-49.
- Lowe, R. K. (1989). "Reading" scientific diagrams: Characterising components of skilled performance. *Research in Science Education*, 18(1), 112-122.
- Lowe, R. K. (1989). Scientific Diagrams: How Well Can Students Read Them? *What Research Says to the Science and Mathematics Teacher* (Volume 3). Perth, Australia: Curtin University of Technology.
- Maier, U., Kleinknecht, M., Metz, K., Schymala, M., & Bohl, T. (2010). Entwicklung und Erprobung eines Kategoriensystems für fächerübergreifende Aufgabenanalyse. *Schulpädagogische Untersuchung Nürnberg*, (Forschungsbericht Nr. 38). Erlangen-Nürnberg: Friedrich-Alexander-Universität.

- Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment*, 1(1), 1-11.
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218.
- Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, 32(1), 1-19.
- Mayer, R. E. (2001). *Multimedia Learning*. Cambridge, New York: Cambridge University Press.
- Mayer, R. E. (2002). Cognitive Theory and the Design of Multimedia Instruction: An Example of the Two-Way Street Between Cognition and Instruction. *New Directions for Teaching & Learning*, 89, 55-71.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43-52
- Mayer, R. E. (2005). Cognitive Theory of Multimedia Learning. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (S. 1-48). New York: Cambridge University Press.
- Mayer, R. E. (2009). *Multimedia Learning. Second Edition*. Cambridge: Cambridge University Press.
- Mayer, R. E. (2014). Cognitive Theory of Multimedia Learning. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (2. Auflagen, S. 43-71). Cambridge: Cambridge University Press.
- Mayer, J., & Wellnitz, N. (2014). Die Entwicklung von Kompetenzstrukturmodellen. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 19-30). Berlin [u.a.]: Springer.
- Mayring, P. (2000). Qualitative Inhaltsanalyse. *Forum Qualitative Sozialforschung (Online Journal)*. Zugriff am 13.06.2014. Verfügbar unter https://www.ph-freiburg.de/fileadmin/dateien/fakultaet3/sozialwissenschaft/Quasus/Volltexte/2-00mayring-d_qualitativeInhaltsanalyse.pdf.
- Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 601-613). Wiesbaden: VS Verlag für Sozialwissenschaften.
- McElvany, N., Schroeder, S., Richter, T., Hachfeld, A., Baumert, J., Schnotz, W., Horz, H., & Ullrich, M. (2010). Texte mit instruktionalen Bildern als Unterrichtsmaterial - Kompetenzen der Lehrkräfte. *Unterrichtswissenschaft*, 38, 98-116.

- Millar, R. (2005). Contextualized science courses: Where next? In P. Nentwig & D. Waddington (Hrsg.), *Making it relevant: Context based learning of science* (S. 246-323). Münster: Waxmann.
- Moosbrugger, H., & Kelava, A. (Hrsg.) (2012). *Testtheorie und Fragebogenkonstruktion. Mit 66 Abbildungen und 41 Tabellen* (Springer Lehrbuch, 2., aktual. Und überarb. Auflage). Berlin [u.a.]: Springer.
- Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 227-274). Berlin: Springer.
- Mortimer, E. F., & Scott, P. H. (2000). Analysing discourse in the science classroom. In J. Leach, R. Millar, & J. Osborne (Hrsg.), *Improving science education: The contribution of research* (S. 126-142). Milton Keynes, UK: Open University Press.
- Muckenfuß, H. (1995). Lernen im sinnstiftenden Kontext. Entwurf einer zeitgemäßen Didaktik des Physikunterrichts (1. Auflage). Berlin: Cornelsen.
- Muckenfuß, H. (2004). Themen und Kontexte als Strukturelemente des naturwissenschaftlichen Unterrichts – Zu den Schwierigkeiten systematisches Physiklernen zu organisieren. *Phy-Did*, 2/3, 57-66.
- Neumann, K. (2014). Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 355-369). Berlin [u.a.]: Springer.
- Neumann, I. (2011). Beyond Physics Content Knowledge. Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge (Studien zum Physik- und Chemielernen, Bd. 117). Berlin: Logos Verlag.
- Neumann, K., Kauertz, A., Lau, A., Notarp, H., & Fischer, H. E. (2007). Die Modellierung physikalischer Kompetenz und ihrer Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften*, 13, 101-121.
- Niebert, K., & Gropengiesser, H. (2015). Understanding Starts in the Mescosm: Conceptual metaphor as a framework for external representations in science teaching. *International Journal of Science Education*, 37(5-6), 903-933.
- Niegemann, H. M., Domagk, S., Hessel, S., Hein, A., Hupfer, M., & Zobel, A. (2008). *Kompendium multimediales Lernen*. Berlin [u.a.]: Springer.
- Nitz, S. (2012). *Fachsprache im Biologieunterricht: Eine Untersuchung zu Bedingungsfaktoren und Auswirkungen*. Dissertation, Christian-Albrechts-Universität zu Kiel, Kiel.
- Nitz, S., Ainsworth, S. E., Nerdel, C., & Prechtel, H. (2014). Do student perceptions of teaching predict the development of representational competence and biological knowledge? *Learning and Instruction*, 31, 13-22.

- Nitz, S., Enzengmüller, C, Precht, H., & Nerdel, C. (2011). Fachsprache im naturwissenschaftlichen Unterricht – eine empirische Untersuchung zur Einstellung angehender Lehrkräfte. *Unterrichtswissenschaft*, 39(3), 245-262.
- Nitz, S., Nerdel, C., & Precht, H. (2012). Entwicklung eines Erhebungsinstrumentes zur Erfassung der Verwendung von Fachsprache im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 18, 117-139.
- Norris, S. P., & Phillips, L. M. (2003). How Literacy in Its Fundamental Sense Is Central to Scientific Literacy. *Science Education*, 87(2), 224-240.
- OECD (2014). *PISA 2012 Technical Report*. Paris: OECD.
- OECD (2009). *PISA 2009 Assessment Framework. Key competencies in reading, mathematics and science*. Paris: OECD.
- OECD (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD.
- Oers, B. van (1998). From context to contextualizing. *Learning and Instruction*, 8(6), 473-488.
- Oestermeier, U. (2008). Lernen mit Text und Bild. Zugriff am 06.07.17. Verfügbar unter https://www.e-teaching.org/didaktik/gestaltung/visualisierung/textbild/Lernen_mit_Text_und_Bild.pdf.
- Oliveira, D. K., Justi, R., Mendonça, P.C. (2015). The Use of Representations and Argumentative and Explanatory Situations. *International Journal of Science Education*, 37(9), 1402-1435.
- Osborne, J. (2002). Science without a literacy: A ship without a sail? *Cambridge Journal of Education*, 32(2), 203-218.
- Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats*. Boston: Kluwer Academic Publishers.
- Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.
- Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Hrsg.), *Artificial intelligence and the future of testing* (S. 73-126). Hillsdale, NJ: Erlbaum.
- Pigge, C., Neumann, I., & Heinze, A. (2016). Mathematische Lernvoraussetzungen für MINT-Studiengänge aus Hochschulsicht – eine Delphi-Studie. In Institut für Mathematik und Informatik der Pädagogischen Hochschule Heidelberg (Hrsg.), *Beiträge zum Mathematikunterricht 2016: Vorträge auf der 50. Tagung für Didaktik der Mathematik vom 07.03.2016 bis 11.03.2016 in Heidelberg* (S. 1501-1502).
- Pohl, S., & Carstensen, C. H. (2012). NEPS Technical Report – Scaling the Data of the Competence Tests. Bamberg: German NEPS.

- Posada, D., & Buckley, T. R. (2004). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biologie*, 53(5), 793-808.
- Pozzer, L. L., & Roth, W.-M. (2003). Prevalance, function, and structure of photographs in high school biology textbooks. *Journal of Research in Science Teaching*, 40(10), 1089-1114.
- Pozzer-Ardenghi, L., & Roth, W.-M. (2010). Toward a Social Practice Perspective on the Work of Reading Incriptions in Science Texts. *Reading Psychology*, 31(3), 228-253.
- Prain, V., & Tytler, R. (2012). Learning through constructing representations in science: A framework of representational construction affordance. *International Journal of Science Education*, 34(17), 2751-2773.
- Prain, V., & Waldrip, B. (2010). Representing science literacy: An introduction. *Research in Science Education*, 40(1), 1-3.
- Prain, V., & Waldrip, B. (2006). An exploratory study of teachers' and students' use of multi-modal representations of concepts in primary science. *International Journal of Science Education*, 28(15), 1843-1866.
- Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, 30(1), 120-135.
- Prenzel, M., Rost, J., Senkbeil, M., Häußler, P., & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 192-250). Opladen: Leske + Budrich.
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (1. Auflage). Studies in Mathematical Psychology I. Kopenhagen: Nielsen & Lydiche.
- Rapp, D. N., & Kurby, C. A. (2008). Mental models: Theoretical issues for visualizations in science education. In J. K. Gilbert, M. Reiner & M. Nakhleh (Hrsg.), *Visualization in Science Education* (S. 43-60). United Kingdom: Springer.
- Rau, M. A., Alevin, V., & Rummel, N. (2015). Successful learning with multiple graphical representations and self-explanation prompts. *Journal of Educational Psychology*, 107(1), 30-46.
- Rauch, D., & Hartig, J. (2008). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. 240-250). Berlin: Springer.
- Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, 52(4), 354-379

- Reckase, M. (2009). *Multidimensional Item Response Theory*. New York: Springer.
- Rincke, K. (2011). It's rather like learning a language: Development of talk and conceptual understanding in mechanics lessons. *International Journal of Science Education*, 33(2), 229-258.
- Rincke, K. (2010). Alltagssprache, Fachsprache und ihre besonderen Bedeutungen für das Lernen. *Zeitschrift für Didaktik der Naturwissenschaften*, 16, 235-260.
- Rincke, K. (2007). *Sprachentwicklung und Fachlernen im Mechanikunterricht* (Bd. 66; H. Niederer, H. Fischler & E. Sumfleth, Hrsg.). Berlin: Logos Verlag.
- Roberts, D. A. (2007). Scientific Literacy/Science Literacy. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of Research on Science Education* S. 729-780). Mahwah: Erlbaum.
- Roelcke, T. (2014). Zur Gliederung von Fachsprache und Fachkommunikation. *Fachsprache – International Journal of Specialized Communication*, 37, 154-178.
- Roelcke, T. (2010). *Fachsprachen* (3., neu bearb. Auflage) (Grundlagen der Germanistik 37). Berlin: Erich Schmidt Verlag.
- Roeling, M. B. (2016). *Konzeptuelles Wissen und Konzeptentwicklung in Biologie* (Biologie lernen und lehren, Bd. 11). Berlin: Logos Verlag.
- Rönnebeck, S., Schöps, K., Prenzel, M., & Hammann, M. (2006). Naturwissenschaftliche Kompetenzen im Ländervergleich. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2006 in Deutschland* (S. 67-92).
- Ropohl, M. (2010). *Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion: Entwicklung und Analyse von Testaufgaben* (Studien zum Physik- und Chemielernen, Bd. 107). Berlin: Logos Verlag.
- Ropohl, M., Walpuski, M., & Sumfleth, E. (2014). Welches Aufgabenformat ist das richtige? – Empirischer Vergleich zweier Aufgabenformate zur standardbasierten Kompetenzmessung. *Zeitschrift für Didaktik der Naturwissenschaften*, 21, 1-15.
- Rost, J., & Davier, M. von (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch Models. Foundations, Recent Developments, and Applications* (S. 257-268). New York: Springer.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (Psychologie Lehrbuch, 2., vollständig überarb. und erw. Auflage). Bern [u.a.]: Hans Huber.
- Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik*, 50(5), 662-678.
- Rost, J. (2006). Kompetenzstrukturen und Kompetenzmessung. *Praxis der Naturwissenschaften – Chemie in der Schule*, 55(8), 5-8.

- Roth, W.-M., & Bowen, G. M. (2001). Professional read graphs: A semiotic analysis. *Journal for Research in Mathematics Education*, 32, 159-194.
- Roth, W.-M., & Pozzer-Ardenghi, L. (2013). Pictures in Biology Education. In D. F. Treagust & C.-Y. Tsui (Hrsg.), *Multiple Representations in Biological Education* (S. 39-53). Dordrecht [u.a.]: Springer.
- Rous, M. (2016). *Fachsprache im Biologieunterricht: Förderung von konzeptuellem Lernen und Textverstehen durch fachspezifisch-sprachsensible Aufgaben* (Biologie lernen und lehren, Bd. 16). Berlin: Logos Verlag.
- Rönnebeck, S., Schöps, K., Prenzel, M., & Hammann, M. (2006). Naturwissenschaftliche Kompetenzen im Ländervergleich. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2006 in Deutschland* (S. 67-92). Münster [u.a.]: Waxmann.
- Rubitzko, T. (2006). Aktives Lernen mit multiplen Repräsentationen – Zur Vermittlung komplexer physikalischer Inhalte mit Texten, Bildern, Animationen und Simulationen. Dissertation. Dissertation, Pädagogische Hochschule Ludwigsburg, Ludwigsburg.
- Rumann, S., Fleischer, J., Stawitz, H. C., Wirth, J., & Leutner, D. (2010). Vergleich von Profilen der Naturwissenschafts- und Problemlöse-Aufgaben der PISA 2003-Studie. *Zeitschrift für Didaktik der Naturwissenschaften*, 16, 315-327.
- Ryken, A. E. (2009). Multiple representations as sites for teacher reflection about mathematics learning. *Journal of Mathematics Teacher Education*, 12: 347.
- Saß, Stefanie (2012). Gestaltungsmerkmale von Testaufgaben in der Kompetenzmessung: empirische Untersuchung zum Einfluss von Text- und Bildelementen auf kognitive Testbearbeitungsprozesse. Dissertation, Christian-Albrechts-Universität zu Kiel, Kiel.
- Schechert, A. (2014). *Entwicklung und Validierung eines Kodiermanuals zur Erfassung von Fähigkeiten im Umgang mit multiplen externen Repräsentationen (MER) im biologischen Kontext*. Masterarbeit, Technische Universität München, München.
- Schecker, H., & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaft*, 12, 45-66.
- Schermelleh-Engel, K., & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119-141). Berlin [u.a.]: Springer.
- Scheiter, K., & Eitel, A. (2015). Signals foster multimedia learning by supporting integration of highlighted text and diagram elements. *Learning and Instruction*, 36, 11-26.
- Scheiter, K., Eitel, A., & Schüler, A. (2016). Lernen mit Texten und Bildern: Die frühzeitige wechselseitige Beeinflussung kognitiver Prozesse bei der Konstruktion eines integrierten mentalen Modells. *Psychologische Rundschau*, 67, 97-103.

- Schlag, S., & Ploetzner, R. (2011). Supporting learning from illustrated texts: conceptualizing and evaluating a learning strategy. *Instructional Science*, 39(6), 921-937.
- Schmiemann, P. (2010). *Modellierung von Schülerkompetenzen im Bereich des biologischen Fachwissens*. Berlin: Logos Verlag.
- Schmiemann, P. (2014). Lernen aus Fehlern. In U. Spörhase & W. Ruppert (Hrsg.), *Biologie Methodik. Handbuch für die Sekundarstufe I und II* (2. Auflage, S. 252-255). Berlin: Cornelsen Schulbuchverlage.
- Schmiemann, P., & Lücken (2014). Validität – Misst mein Test, was er soll? In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 107-118). Berlin [u.a.]: Springer.
- Schnotz, W. (1994). Wissenserwerb mit logischen Bildern. In B. Weidenmann (Hrsg.), *Wissenserwerb mit Bildern: Instruktionale Bilder in Printmedien, Film/Video und Computerprogrammen* (S. 95-148). Bern: Huber.
- Schnotz, W. (2001). Sign systems, technologies, and the acquisition of knowledge. In J.-F. Rouet, J. Levonen & A. Biarreau (Eds.), *Multimedia Learning: Cognitive and Instructional Issues* (pp. 9-29). Amsterdam, The Netherlands: Pergamon.
- Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review*, 14(1), 101-120.
- Schnotz, W. (2003). Bild- und Sprachverarbeitung aus psychologischer Sicht. In K. Sachs-Hombach (Hrsg.), *Was ist Bildungskompetenz? Studien zur Bildwissenschaft* (S. 25-42). Wiesbaden: Deutscher Universitäts-Verlag.
- Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (S. 49-70). New York: Cambridge University Press.
- Schnotz, W. (2010). Visuelles Lernen. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 927-935).
- Schnotz, W., & Bannert, M. (1999). Einflüsse der Visualisierungsform auf die Konstruktion mentaler Modelle beim Text- und Bildverstehen. *Zeitschrift für Experimentelle Psychologie*, 46(3), 217-236.
- Schnotz, W., Bannert, M., & Seufert, T. (2002). Toward an integrative view of text and picture comprehension: Visualization effects on the construction of mental models. In J. Otero, J. A. León & A. C. Graesser (Hrsg.), *The psychology of science text comprehension* (S. 385-416). Mahwah: Erlbaum.
- Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13(2), 141-156.

- Schnotz, W., Ullrich, M., Hochpöchler, U., Horz, H., McElvany, N., Schroeder, S., & Baumert, J. (2011). What makes text-picture-integration difficult? A structural and procedural analysis of textbook requirements. *Ricerche di Psicologia*, 1, 103-135.
- Schnotz, W., Horz, H., McElvany, N., Schroeder, S., Ullrich, M., Baumert, J., Hachfeld, A., & Richter, T. (2010). Das BITE-Projekt: Integrative Verarbeitung von Bildern und Texten in der Sekundarstufe. Projekt BITE. In E. Klieme, D. Leutner, M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (S. 143-153). Weinheim, Basel: Beltz.
- Schnotz, W., Picard, E., & Hron, a. (1993). *How do successful and unsuccessful learners use texts and graphics?* Tübingen: Universität, Deutsches Institut für Fernstudien, Arbeitsbereich Lernforschung
- Schönborn, K. J., & Anderson, T. R. (2009). A model of factors determining students' ability to interpret external representations in biochemistry. *International Journal of Science Education*, 31(2), 193-232.
- Schönborn, K. J., & Anderson, T. R. (2008). Bridging the educational research-teaching practice gap. *Biochemistry and Molecular Biology Education*, 36(4), 309-315.
- Schönborn, K. J., & Bögeholz, S. (2013). Experts' Views on Translation Across Multiple External Representations in Acquiring Biological Knowledge About Ecology, Genetics, and Evolution. In D. F. Treagust & C.-Y. Tsui (Hrsg.), *Multiple Representations in Biological Education* (S. 111-128). Dordrecht [u.a.]: Springer Science+Business Media B.V.
- Schönborn, K. J., & Bögeholz, S. (2009). Knowledge transfer in Biology and Translation across external Representations: Experts' Views and Challenges for Learning. *International Journal of Science and Mathematics Education*, 7(5), 931-955.
- Schrader, F.-W., & Helmke, A. (2002). Alltägliche Leistungsbeurteilung durch den Lehrer. In F.-E. Weinert (Hrsg), *Leistungsmessungen in Schulen* (S. 45-58). Weinheim: Beltz.
- Schüler, A., Arndt, J., & Scheiter, K. (2015). Processing multimedia material: Does integration of text and picture result in a single or two interconnected mental representations? *Learning and Instruction*, 35, 62-72.
- Schütte, K., Frenzel, A. C., Asseburg, R., & Pekrun, R. (2007). Schülermerkmale, naturwissenschaftliche Kompetenz und Berufserwartung. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 125-146). Münster: Waxmann.
- Schwonke, R., Berthold, K., & Renkl, A. (2009). How multiple external representations are used and how they can be made more useful. *Applied Cognitive Psychology*, 23(9), 1227-1243.
- Sedlmeier, P., & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie (Person Studium – Psychologie)*. München: Pearson.

- Seidel, T., Prenzel, M., Wittwer, J., & Schwindt, K. (2007). Unterricht in den Naturwissenschaften. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 147-179). Münster: Waxmann.
- Seidel, T., Prenzel, M., Rimmel, R., Dalehefte, I. M., Herweg, C., & Kobarg, M. et al. (2006). Blicke auf den Physikunterricht: Ergebnisse der IPN Videostudie. *Zeitschrift für Pädagogik*, 52, 799-821.
- Senkbeil, M., Rost, J., Carstensen, C. H., & Walter, O. (2005). Der nationale Naturwissenschaftstest PISA 2003. Entwicklung und empirische Überprüfung eines zweidimensionalen Facettendesigns. *Empirische Pädagogik*, 19(2), 166-189.
- Seufert, T. (2003). Supporting coherence from formation in learning from multiple representations. *Learning and Instruction*, 13(2), 227-237.
- Shavelson, R. J. (2013). An approach to testing & modeling competence. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Hrsg.), *Modeling and measuring competencies in higher education. Tasks and Challenges* (S. 29-43). Rotterdam: Sense Publishers.
- Spoden, C., Fleischer, J., & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik*, 35(1), 79-99.
- Stanat, P., & Kunter, M. (2001). Geschlechterunterschiede in Basiskompetenzen. In Deutsches PISA-Konsortium (Hrsg.), S. 249-269.
- Stäudel, L., Franke-Braun, G., & Parchmann, I. (2008). Sprache, Kommunikation und Wissenserwerb im Chemieunterricht. *Naturwissenschaften im Unterricht Chemie*, 19(106/107), 4-9.
- Stecher, B. M., & Klein, S. P. (1997). The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educational Evaluation and Policy Analysis*, 19(1), 1-14.
- Stenning, K., & Oberlander, J. (1995). A Cognitive Theory of Graphical and Linguistic Reasoning: Logic and Implementation. *Cognitive Science*, 19(1), 97-140.
- Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction*, 13, 191-203.
- Stieff, M. (2011). Improving representational competence using molecular simulations embedded in inquiry activities. *Journal of Research in Science Teaching*, 48(10), 1137-1158.
- Strobl, C. (2012). Das Rasch-Modell – Eine verständliche Einführung für Studium und Praxis (2. Erw. Auflage). München [u.a.]: Rainer Hampp Verlag.

- Sumfleth, E., & Pitton, A. (1998). Sprachliche Kommunikation im Chemieunterricht: Schüler- vorstellungen und ihre Bedeutung im Unterrichtsalltag. *Zeitschrift für Didaktik der Naturwissenschaften*, 4(2), 4-20.
- Suwelack, W. (2010). Lehren und Lernen im kompetenzorientierten Unterricht. Modellvorstellungen für die Praxis: Vom Kompetenzmodell zum Prozessmodell. *Der Mathematische und Naturwissenschaftliche Unterricht*, 63(1), 176-182.
- Sweller, J., & Chandler, P. A. (1994). Why some material is difficult to learn. *Cognition and Instruction*, 12(3), 185-233.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (S. 19-30). New York: Cambridge University Press.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3), 251-196.
- Sweller, M. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295-312.
- Tabbers, H. K. (2002). *The modality of text in multimedia instructions: Refining the design guidelines*. Maastricht, The Netherlands: Datawyse Maastricht.
- Terzer, E. (2012). *Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice Items*. Dissertation, Humboldt-Universität zu Berlin. Berlin.
- Terzer, E., Hartig, J., & Upmeyer zu Belzen, A. (2013). Systematische Konstruktion eines Tests zu Modellkompetenz im Biologieunterricht unter Berücksichtigung von Gütekriterien. *Zeitschrift für Didaktik der Naturwissenschaften*, 19, 51-76.
- Theyßen, H., Schmidt, M., Einhaus, E., & Schecker, H. (2006). Ein indikatorenbasiertes Verfahren zur Einstufung von Testaufgaben in ein Kompetenzmodell. *Physik und Didaktik in Schule und Hochschule*, 2/5, 123-134.
- Treagust, D. F., & Tsui, C.-Y. (2013). Conclusion: Contributions of Multiple Representations to Biological Education. In D. F. Treagust & E. Tsui (Hrsg.), *Multiple Representations in Biological Education* (S. 349-368). Dordrecht: Springer.
- Treagust, D. F., & Tsui, C.-Y. (2013). Contributions of Multiple Representations to Biological Education. *Paper presented at ASERA Conference 2013*. New Zealand: Wellington.
- Tsui, C.-Y., & Treagust, D. F. (2003). Genetics Reasoning with Multiple External Representations. *Research in Science Education*, 33, 111-135.
- Tsui, C.-Y., & Treagust, D. F. (2013). Introduction to Multiple Representations: Their Importance in Biology and Biological Education. In D. F. Treagust & E. Tsui (Hrsg.), *Multiple Representations in Biological Education* (S. 3-18). Dordrecht: Springer.

- Ummels, M. H. J., Kamp, M. J. A., Kroon, H. de, & Boersma, K. T. (2015). Designing and Evaluating a Context-based Lesson Sequence Promoting Conceptual Coherence in Biology. *Journal of Biological Education*, 49(1), 38-52.
- Unterbruner, U. (2007). Multimedia-Lernen und Cognitive Load. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 153-164). Berlin [u.a.]: Springer.
- Van de Walle, S. (2004). Context-specific images of the archetypical bureaucrat: persistence and diffusion of the bureaucracy stereotype. *Public Voices*, 7(1), 3-12.
- Vock, K. (2009). Eine experimentelle Studie zur Itemschwierigkeit von Freiem Antwortformat sowie "1 aus 4-", "1 aus 6"- und "x aus 5"-MC-Antwortformat. Diplomarbeit, Universität Wien, Wien.
- Vorst, H. van, Dorschu, A., Fechner, S., Kauertz, A., Krabbe, H., & Sumfleth, E. (2014). Charakterisierung und Strukturierung von Kontexten im naturwissenschaftlichen Unterricht – Vorschlag einer theoretischen Modellierung. *Zeitschrift für Didaktik der Naturwissenschaften*, 21, 29-39.
- Wafi, S., & Wirtz, M. A. (2016). *Visualisierungskompetenz in didaktischen Kontexten. Eine Einführung*. Wiesbaden: Springer VS.
- Waldrup, B. G., Prain, V., & Carolan, J. (2010). Using multi-modal representations to improve learning in junior secondary science. *Research in Science Education*, 40(1), 65-80.
- Walzer, M., Fischer, H. E., & Borowski, A. (2014). Fachwissen im Studium zum Lehramt der Physik. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht* (S. 108-110). Kiel: IPN.
- Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychological Methods*, 9(1), 116-136.
- Watzka, B., & Girwidz, R. (2015). Einfluss der Kontextorientierung und des Präsentationsmodus von Aufgaben auf den Wissenserwerb und die Transferleistung physikalischer Inhalte. *Zeitschrift für Didaktik der Naturwissenschaften*, 21, 187-206.
- Webb, N. L. (2006). Identifying content for student achievement tests. In S. Downing & T. Haladyna (Hrsg.), *Handbook of Test Development* (S. 299-309). New Jersey: Lawrence Erlbaum.
- Weidenmann, B. (1994). *Wissenserwerb mit Bildern. Instruktionale Bilder in Printmedien, Film/Video und Computerprogrammen* (1. Auflage). Bern [u.a.]: Hans Huber.
- Weidenmann, B. (2002). Multicodierung und Multimodalität im Lernprozess. In L. J. Issing & P. Klimsa (Hrsg.), *Information und Lernen mit Multimedia und Internet* (Vol. 3, S. 45-64). Weinheim: Beltz.

- Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schule – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17-31). Weinheim [u.a.]: Beltz.
- Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. Buckingham, PA: Open University Press.
- Wellnitz, N. (2012). *Kompetenzstruktur und –niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung* (Biologie lernen und lehren, Bd. 2). Berlin: Logos Verlag.
- Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A., Sumfleth, E., & Walpuski, M. (2012). Evaluation der Bildungsstandards – eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften*, 18, 261-291.
- Werner, M., Schwanewedel, J., & Mayer, J. (2014). Does the context make a difference? Students' abilities in decision-making and the influence of contexts. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *Science Education Research For Evidence-based Teaching and Coherence in Learning. Part 8 Scientific Literacy and socio scientific issues* (S. 81-89). Nicosia: European Science Education Research Association.
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, 66, 279-333.
- Williams, G., & Clement, J. (2015). Identifying Multiple Levels of Discussion-Based Teaching Strategies for Constructing Scientific Models. *International Journal of Science Education*, 37(1), 82-107.
- Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. New York: Psychology Press.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370-371.
- Wright, B. D., & Stone, M. H. (1999). *Measurement Essentials. 2nd Edition*. Wilmington [u.a.]: Wide Range.
- Wu, H.-K., Kuo, C.-Y., Jen, T.-H., & Hsu, Y.-S. (2015). What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers & Education*, 85, 35-48.
- Wu, H.-K., & Puntambekar, S. (2012). Pedagogical affordance of multiple external representations in scientific processes. *Journal of Science Education and Technology*, 21(6), 754-767.
- Wu, M. L. (2005). The Role of Plausible Values in Large-Scale Surveys. *Studies in Educational Evaluation*, 31, 114-128.

- Wu, M. L., & Adams, R. (2007). Applying the Rasch Model to Psycho-Social Measurement: A Practical Approach. Educational Measurement Solutions. Zugriff am 29.06.2017. Verfügbar unter https://media.metrik.de/uploads/incoming/pub/Literatur/von%20Winfried/RaschMeasurement_Complete.pdf.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0. Generalised Item Response Modelling Software*. Camberwell: ACER Press.
- Yore, L. D., & Hand, B. (2010). Epilogue: Plotting a research agenda for multiple representations, multiple modality, and multimodal representational competency. *Research in Science Education*, 40(1), 93-101.
- Yore, L. D., Primm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science and Mathematics Education*, 5(4), 559-589.
- Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, 25(6), 689-725.
- Ziepprecht, K. (2016). *Strategien und Kompetenzen von Lernenden beim Erschließen von biologischen Informationen aus unterschiedlichen Repräsentationen* (Biologie lernen und lehren, Bd. 15). Berlin: Logos Verlag.
- Zimbardo, P. G., & Gerrig, R. J. (1999). *Psychologie* (7. Auflage). Berlin: Springer.
- Zöfel, P. (2003). *Statistik für Psychologen. Im Klartext*. München: Pearson Studium.

13.2 TABELLENVERZEICHNIS

TABELLE 1: VORSTUDIE - STUDIENGÄNGE UND ANZAHL DER STUDIERENDEN (N=730).....	68
TABELLE 2: VORSTUDIE - ITEMS IN UNTERSCHIEDLICHEN ANTWORTFORMATEN	81
TABELLE 3: ITEMVERTEILUNG VORSTUDIE (N=210).....	82
TABELLE 4: VORSTUDIE - TESTHEFTDESIGN ANGELEGT ALS YODEN-SQUARE-DESIGN (210 ITEMS)	84
TABELLE 5: CODES FÜR DIE ITEMSPEZIFISCHE KODIERUNG	94
TABELLE 6: RICHTWERT FÜR COHENS KAPPA KOEFFIZIENT	97
TABELLE 7: KREUZTABELLE ZUR BERECHNUNG DER INTRATER-RELIABILITÄT	98
TABELLE 8: KREUZTABELLE ZUR BERECHNUNG DER INTERRATER-RELIABILITÄT	98
TABELLE 9: DEDUKTIVE KATEGORIENBILDUNG AUF BASIS DER THEORIEBASIERTEN INHALTSANALYSE	107
TABELLE 10: DEDUKTIVE FÄHIGKEITEN UND ANFORDERUNGSBEREICHE IM UMGANG MIT MER....	108
TABELLE 11: ERWEITERUNG ZU EINER INDUKTIV-DEDUKTIVEN KATEGORIENBILDUNG (MAYRING, 2000)	109
TABELLE 12: KATEGORIENSYSTEM ZUR CODIERUNG VON OFFENEN AUFGABEN MIT MER	110
TABELLE 13: ITEMKENNWERTE DER EINDIMENSIONALEN FALLZENTRIERTEN RASCH-SKALIERUNG VOR SELEKTION (95 ITEMS DER A-SERIE).....	114
TABELLE 14: ITEMKENNWERTE DER EINDIMENSIONALEN FALLZENTRIERTEN RASCH-SKALIERUNG NACH SELEKTION (55 ITEMS DER A-SERIE).....	114
TABELLE 15: ITEMSELEKTION AUFGRUND DER RASCHANALYSE (N=55 ITEMS)	119
TABELLE 16: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER MER-INTEGRATION, GETRENNT NACH AUFGABENKONTEXTEN UND DEM REPRÄSENTATIONSTYP (N=128 ITEMS).....	121
TABELLE 17: HAUPTSTUDIE - TEILGENOMMENE STUDIENGÄNGE UND ANZAHL DER STUDIERENDEN (N=548).....	128
TABELLE 18: HAUPTSTUDIE- VERTEILUNG DER TESTAUFGABEN (ITEMS) FÜR DAS KOMPETENZKONSTRUKT "UMGANG MIT MER" IN DREI AUFGABENKONTEXTEN UND ZWEI REPRÄSENTATIONSTYPEN (N=36 ITEMS)	130
TABELLE 19: HAUPTSTUDIE - VERTEILUNG DER ITEMS AUF DIE MER-TEILKOMPONENTEN (I, II, III) UNTER BERÜCKSICHTIGUNG VON AUFGABENKONTEXT UND REPRÄSENTATIONSTYP	131
TABELLE 20: HAUPTSTUDIE – FINALE ITEMSAUSWAHL, GETRENNT NACH MER-INTEGRATION, AUFGABENKONTEXT UND REPRÄSENTATIONSTYP	131
TABELLE 21: HAUPTSTUDIE – ITEMS IN UNTERSCHIEDLICHEN ANTWORTFORMATEN	132
TABELLE 22: HAUPTSTUDIE – TESTHEFTDESIGN (36 ITEMS)	133

TABELLE 23: EINSTUFUNG DES KORRELATIONSKOEFFIZIENTEN R (NACH ZÖFEL, 2003)	147
TABELLE 24: KENNWERTE DER EINDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 1-DIM-MODELL (N=548)	153
TABELLE 25: KENNWERTE DER DREIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 3-DIM MODELL DER MER-INTEGRATION.....	156
TABELLE 26: KENNWERTE DER DREIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 3DIM-BETWEEN-MODELL DER AUFGABENKONTEXTE ÖKOLOGIE, STOFFWECHSEL, GENETIK.....	159
TABELLE 27: KENNWERTE DER ZWEIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 2DIM-BETWEEN-MODELL DER REPRÄSENTATIONSTYPEN DIAGRAMM UND SCHEMA	161
TABELLE 28: KENNWERTE DER ZWEIDIMENSIONALEN, FALLZENTRIERTEN RASCH-SKALIERUNG FÜR DAS 2-DIM-BETWEEN-MODELL DER MER-INTEGRATION	163
TABELLE 29: KENNWERTE DER ITEMZENTRIERTEN RASCH-SKALIERUNGEN FÜR DIE EIN- UND MEHRDIMENSIONALEN MODELLIERUNGEN	166
TABELLE 30: MODELLPRÜFUNG NACH DEM AIC, BIC UND CAIC.....	167
TABELLE 31: LATENTE KORRELATION (ϕ) ZWISCHEN DEN DIMENSIONEN MERI UND MERII (ϕ) BZW. KORRELATION DER PERSONENPARAMETER	168
TABELLE 32: LATENTE KORRELATIONEN (ϕ) ZWISCHEN DEN DIMENSIONEN MERI, MERII, MERII BZW. KORRELATION DER PERSONENPARAMETER	168
TABELLE 33: LATENTE KORRELATIONEN (ϕ) ZWISCHEN DEN DIMENSIONEN ÖKO, STW, GEN BZW. KORRELATION DER PERSONENPARAMETER	169
TABELLE 34: LATENTE KORRELATIONEN (ϕ) ZWISCHEN DEN DIMENSIONEN MERI, MERII, MERIII UND DEN AUFGABENKONTEXTEN (ÖKO, STW, GEN) FÜR DAS 6D-WITHIN-MODELL BZW. KORRELATION DER PERSONENPARAMETER	169
TABELLE 35: LINEARE REGRESSION ZUR ABHÄNGIGKEIT DER SCHWIERIGKEIT VON DER KOMPONENTE MER-INTEGRATION (ALS DUMMY-VARIABLE IN DAS MODELL EINBEZOGEN) .	173
TABELLE 36: VERGLEICH DER MITTLEREN ITEMSCHWIERIGKEITEN NACH KOMPONENTEN DER MER- INTEGRATION, GETRENNT NACH AUFGABENKONTEXTEN UND DEM REPRÄSENTATIONSTYP (N=36 ITEMS).....	176
TABELLE 37: MULTIPLE LINEARE REGRESSION ZUR ABHÄNGIGKEIT DER SCHWIERIGKEIT VON DEN DIMENSIONEN MER-INTEGRATION MIT DEN KOMPONENTEN MERI UND MERII SOWIE VON DER DIMENSION AUFGABENKONTEXT MIT DEN KOMPONENTEN ÖKOLOGIE, STOFFWECHSEL UND GENETIK (ALS DUMMY-VARIABLE IN DAS MODELL EINBEZOGEN).....	178
TABELLE 38: VERGLEICH DER STUDIERENDENFÄHIGKEITEN IN DEN KOMPETENZEN MERI UND MERII DIFFERENZIERT NACH STUDIENGANG (N=533).....	182

TABELLE 39: MULTIPLE LINEARE REGRESSION ZUR ABHÄNGIGKEIT DER PERSONENFÄHIGKEIT VON DER KOMPONENTE STUDIENGANG (ALS DUMMY-VARIABLEN IN DAS MODELL EINBEZOGEN)	185
TABELLE 40: ANTWORTBEISPIELE ZUM ITEM FLO08 ZU DEN TEILKOMPONENTEN MERI, ÖKOLOGIE, DIAGRAMM (VGL. MUSTERLÖSUNG IM ANHANG 14.1)	187
TABELLE 41: ANTWORTBEISPIELE ZUM ITEM Dna02 ZU DEN TEILKOMPONENTEN MERI, GENETIK, SCHEMA (VGL. MUSTERLÖSUNG IM ANHANG 14.1)	188

13.3 ABBILDUNGSVERZEICHNIS

ABBILDUNG 1: EINTEILUNG VON BILDERN (ANGELEHNT AN SCHNOTZ, 1994).....	16
ABBILDUNG 2: STRUKTURMODELL FÜR FÄHIGKEITEN BEIM DIAGRAMMGEBRAUCH (LACHMAYER, NERDEL, & PRECHTL, 2007, S. 156, TAB. 1; LACHMAYER, 2008)	21
ABBILDUNG 3: 'THE CUBE MODEL' FOR LEARNING BIOLOGY WITH MER, VERÄNDERT NACH TSUI & TREAGUST, 2013, S. 10, FIG. 1.2	37
ABBILDUNG 4: MEHRSPEICHERMODELL NACH ATKINSON & SHIFFRIN (1968), VERÄNDERT NACH ATKINSON & SHIFFRIN, 1971, S. 3B, FIGURE 1	39
ABBILDUNG 5: KOGNITIVE THEORIE DES MULTIMEDIALERNEN NACH MAYER (1997), AUF BASIS VON MAYER, 2009, S. 61, FIGURE 3.1.	41
ABBILDUNG 6: INTEGRATIVES MODELL DES TEXT-BILD-VERSTEHENS NACH SCHNOTZ & BANNERT, 1999, AUF BASIS VON SCHNOTZ, 2001, S. 306, ABBILDUNG 1	44
ABBILDUNG 7: KOMPETENZSTRUKTURMODELL ZUM UMGANG MIT MULTIPLLEN EXTERNEN REPRÄSENTATIONEN (MER) IM BIOLOGISCHEN KONTEXT	55
ABBILDUNG 8: BEISPIEL FÜR ZWEI ITEMS ZUM DIAGRAMMSPEIFISCHEN VORWISSEN	69
ABBILDUNG 9: SCHRITTE DER TEST- UND ITEMKONSTRUKTION UND DURCHFÜHRUNG DER VOR- UND HAUPTSTUDIE	70
ABBILDUNG 10: MODELLKONFORMES AUFGABENBEISPIEL ZUM UMGANG MIT MER, TEILKOMPETENZ <i>INFORMATIONSENTNAHMENAHE INTEGRATION (MERI)</i>	75
ABBILDUNG 11: MODELLKONFORMES AUFGABENBEISPIEL ZUM UMGANG MIT MER, TEILKOMPETENZ <i>KONSTRUKTIONSNÄHE INTEGRATION (MERII)</i>	77
ABBILDUNG 12: MODELLKONFORMES AUFGABENBEISPIEL ZUM UMGANG MIT MER, TEILKOMPETENZ <i>TRANSLATION UND TRANSFORMATION (MERIII)</i>	78
ABBILDUNG 13: ZUSAMMENHANG VON PERSONENMERKMAL UND TESTVERHALTEN (IN ANLEHNUNG AN ROST, 2004, S. 21)	88
ABBILDUNG 14: LATENTE VARIABLE UND MANIFESTE MERKMALSAUSPRÄGUNGEN	88
ABBILDUNG 15: 1-DIMENSIONALES MODELL ZUR INTEGRATION VON MER.....	91
ABBILDUNG 16: DIE ITEMCHARAKTERISTIK-KURVE (ICC) DES RASCH-MODELLS FÜR EIN ITEM MIT DER ITEMSCHWIERIGKEIT $\sigma_i=0$	99
ABBILDUNG 17: EINDIMENSIONALE FALLZENTRIERTE RASCH-SKALIERUNG VOR SELEKTION (95 ITEMS DER A-SERIE; EIN X STEHT FÜR 0,6 STUDIERENDE).....	113
ABBILDUNG 18: EINDIMENSIONALE FALLZENTRIERTE RASCH-SKALIERUNG NACH SELEKTION (55 ITEMS DER A-SERIE, EIN X STEHT FÜR 0,6 STUDIERENDE).....	115
ABBILDUNG 19: EINDIMENSIONALE FALLZENTRIERTE RASCH-SKALIERUNG NACH SELEKTION (73 ITEMS DER B-SERIE, EIN X STEHT FÜR 0,5 STUDIERENDE).....	117

ABBILDUNG 20: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER MER-INTEGRATION FÜR DIE DREI AUFGABENKONTEXTE (N=128 ITEMS, M = MITTELWERT)	120
ABBILDUNG 21: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER MER-INTEGRATION FÜR DIE ZWEI REPRÄSENTATIONSTYPEN (N=128 ITEMS, M = MITTELWERT).....	123
ABBILDUNG 22: MITTLERE AUFGABENSCHWIERIGKEITEN (ITEMPARAMETER) NACH KOMPONENTEN DER AUFGABENKONTEXTE FÜR DIE ZWEI REPRÄSENTATIONSTYPEN DIAGRAMM UND SCHEMA (N=128 ITEMS, M = MITTELWERT).....	124
ABBILDUNG 23: SCHEMATISCHE DARSTELLUNG DER MODELLIERUNG VON TESTAUFGABEN X UND Y MIT A) ZWEIDIMENSIONALES IRT-MODELL MIT BETWEEN-ITEM-STRUKTUR UND B) ZWEIDIMENSIONALES IRT-MODELL MIT WITHIN-ITEM-STRUKTUR (ANGELEHNT AN HARTIG & HÖHLER, 2010)	136
ABBILDUNG 24: 3-DIMENSIONALES MODELL (INTEGRATION VON MER ALS MERI, MERII, MERIII), BETWEEN-ITEM-MEHRDIMENSIONALITÄT UND ZU PRÜFENDER KORRELATIONEN ZWISCHEN DEN LATENTEN VARIABLEN	139
ABBILDUNG 25: 1-DIMENSIONALES MODELL ZUR INTEGRATION VON MER	139
ABBILDUNG 26: 3-DIMENSIONALES (INTEGRATION NACH AUFGABENKONTEXT) UND 2-DIMENSIONALES MODELL (INTEGRATION NACH REPRÄSENTATIONSTYP), BETWEEN-ITEM-MEHRDIMENSIONALITÄT	140
ABBILDUNG 27: MEHRDIMENSIONALES MODELL MIT UNTERSCHIEDLICHEN KOMBINATIONSMÖGLICHKEITEN ZUR WITHIN-ITEM-MEHRDIMENSIONALITÄT	140
ABBILDUNG 28: 2-DIMENSIONALES MODELL ZUR INTEGRATION VON MER (MERI, MERII), BETWEEN-ITEM-MEHRDIMENSIONALITÄT	142
ABBILDUNG 29: EINDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG ZUR INTEGRATION VON MER (EIN X REPRÄSENTIERT 0,8 STUDIERENDE).....	154
ABBILDUNG 30: DREIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 3-DIM-MODELL DER MER-INTEGRATION (EIN X STEHT FÜR 3,2 STUDIERENDE).....	155
ABBILDUNG 31: DREIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 3-DIM-BETWEEN-MODELL DER AUFGABENKONTEXTE (EIN X STEHT FÜR 4,3 STUDIERENDE).....	158
ABBILDUNG 32: ZWEIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 2-DIM-BETWEEN-MODELL REPRÄSENTATIONSTYP DIAGRAMM UND SCHEMA (EIN X STEHT FÜR 2,1 STUDIERENDE)	160
ABBILDUNG 33: ZWEIDIMENSIONALE, FALLZENTRIERTE RASCH-SKALIERUNG, 2-DIM-BETWEEN-MODELL DER MER-INTEGRATION (EIN X STEHT FÜR 1,9 STUDIERENDE)	162

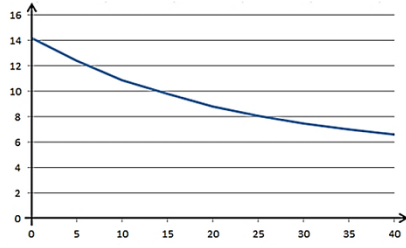
ABBILDUNG 34: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH KOMPONENTEN DER MER- INTEGRATION (DREIDIMENSIONAL, N=36 ITEMS, M = MITTELWERT)	172
ABBILDUNG 35: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH KOMPONENTEN DER MER- INTEGRATION (ZWEIDIMENSIONAL, N=36 ITEMS, M = MITTELWERT).....	172
ABBILDUNG 36: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH AUFGABENKONTEXT (N=36 ITEMS, M = MITTELWERT).....	174
ABBILDUNG 37: VERTEILUNG DER ITEMSCHWIERIGKEIT NACH REPRÄSENTATIONSTYP (N=36 ITEMS)	175
ABBILDUNG 38: PERSONENFÄHIGKEITEN BEI DER MER-INTEGRATION MIT MERI: INFORMATIONSENTNAHMENAEHE INTEGRATION UND MERII: KONSTRUKTIONSAHE INTEGRATION (M = MITTELWERT)	180
ABBILDUNG 39: PERSONENFÄHIGKEITEN BEIM REPRÄSENTATIONSTYP DIAGRAMM UND SCHEMA (M = MITTELWERT)	180
ABBILDUNG 40: PERSONENFÄHIGKEITEN BEI DEN AUFGABENKONTEXTEN ÖKOLOGIE, STOFFWECHSEL, GENETIK (M = MITTELWERT).....	181
ABBILDUNG 41: PERSONENFÄHIGKEITEN DER MERI-INTEGRATION NACH STUDIENGANG (N=533, M = MITTELWERT)	183
ABBILDUNG 42: PERSONENFÄHIGKEITEN DER MERII-INTEGRATION NACH STUDIENGANG (N=533, M = MITTELWERT)	183
ABBILDUNG 43: ANTWORTBEISPIELE ZUM ITEM SEE07 ZU DEN TEILKOMPONENTEN MERII, ÖKOLOGIE, DIAGRAMM (VGL. MUSTERLÖSUNG IM ANHANG 14.1.1).....	188
ABBILDUNG 44: ANTWORTBEISPIELE ZUM ITEM EPO05 - WERTE ENTNEHMEN UND EINTRAGEN (MERII, ÖKOLOGIE, DIAGRAMM)	189
ABBILDUNG 45: ANTWORTBEISPIELE ZUM ITEM EPO05 – DIAGRAMMTYP	190
ABBILDUNG 46: ANTWORTBEISPIELE ZUM ITEM EPO05 – ACHSENZUORDNUNG UND SKALIERUNG	190
ABBILDUNG 47: ANTWORTBEISPIELE ZUM ITEM BLU05 ZU DEN TEILKOMPONENTEN MERII, GENETIK, SCHEMA (VGL. MUSTERLÖSUNG IM ANHANG 14.1.1)	191
ABBILDUNG 48: ANTWORTBEISPIELE ZUM ITEM SEE09 - SKALIERUNG DER ZIELVARIABLE (MERIII, ÖKOLOGIE, DIAGRAMM)	192
ABBILDUNG 49: ANTWORTBEISPIEL ZUM ITEM SEE09 - VERLAUF DER ZIELVARIABLE	193
ABBILDUNG 50: ANTWORTBEISPIELE (A, B, C) ZUM ITEM ERB05 (MERIII, GENETIK, SCHEMA, VGL. MUSTERLÖSUNG ANHANG 14.1.1, S. 267)	194
ABBILDUNG 51: ANTWORTBEISPIEL ZUM ITEM ERB05	194

ANHANG

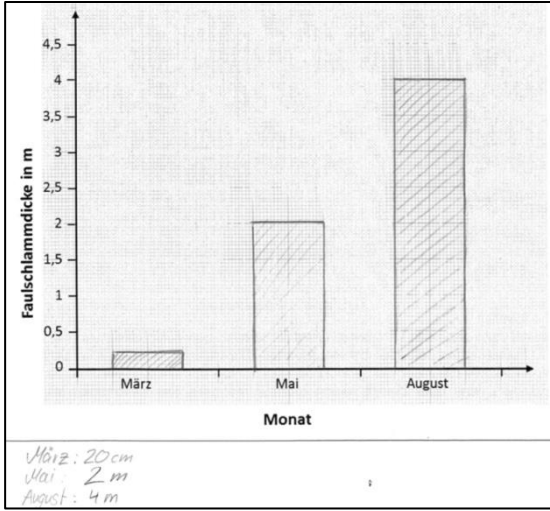
14. 1 TESTINSTRUMENTE

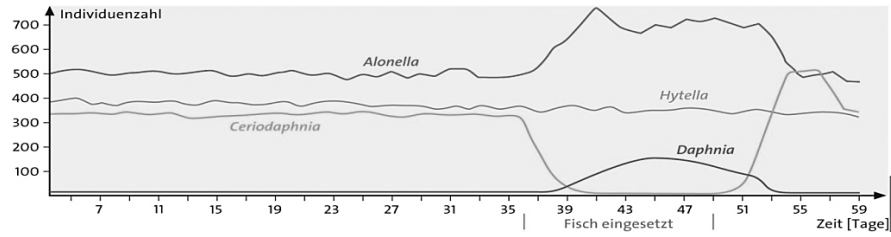
14.1.1 ITEMBEISPIELE DER HAUPTSTUDIE

Kontext Ökologie

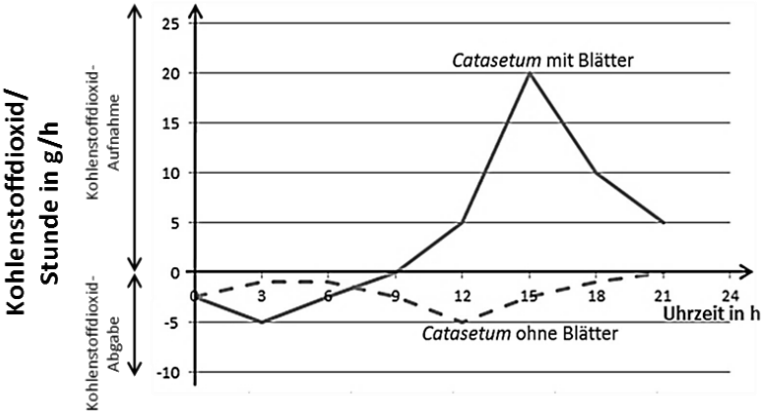
Item See02	Informationsent- nahmenahme Integration (MERI)	Ökologie	Diagramm																				
Quelle	NATURA Biologie für Gymnasien, Oberstufe Ernst Klett Verlag, 2010																						
Aufgabentitel	Das Ökosystem See																						
Aufgaben- stamm und Im- puls	Bei einem Freilandpraktikum im Limnologischen Institut für Gewässer- forschung sieht sich Paul die Aufzeichnungen über den Herrensee an.																						
	Paul weiß, dass Sauerstoff im Was- ser als Gas gelöst ist und findet das nebenstehende Diagramm. Es zeigt den Sauerstoffgehalt des Sees in mg O ₂ /l in Abhängigkeit von der Temperatur.		 <table border="1"> <caption>Data points from the graph</caption> <thead> <tr> <th>Temperature (°C)</th> <th>Oxygen Content (mg O₂/l)</th> </tr> </thead> <tbody> <tr><td>0</td><td>14</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>10</td><td>10</td></tr> <tr><td>15</td><td>9</td></tr> <tr><td>20</td><td>8</td></tr> <tr><td>25</td><td>7.5</td></tr> <tr><td>30</td><td>7</td></tr> <tr><td>35</td><td>6.5</td></tr> <tr><td>40</td><td>6</td></tr> </tbody> </table>	Temperature (°C)	Oxygen Content (mg O ₂ /l)	0	14	5	12	10	10	15	9	20	8	25	7.5	30	7	35	6.5	40	6
Temperature (°C)	Oxygen Content (mg O ₂ /l)																						
0	14																						
5	12																						
10	10																						
15	9																						
20	8																						
25	7.5																						
30	7																						
35	6.5																						
40	6																						
Fragestellung	Beschreiben Sie zunächst den Kurvenverlauf und vergleichen Sie an- schließend die beiden Temperaturintervalle 0-10 °C und 20-30 °C. Begründen Sie damit, wie sich der Sauerstoffgehalt in Abhängigkeit von der Temperatur im Sommer und Winter verhält.																						
Antwort	_____																						

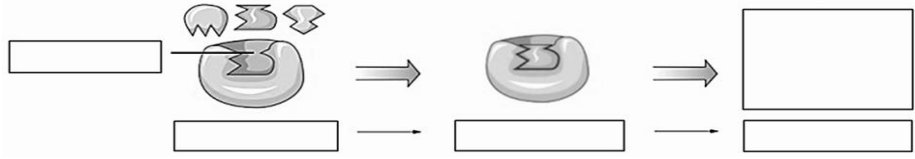
Item See07	Konstruktionsnahe Integration (MERII)	Ökologie	Diagramm														
Quelle	NATURA Biologie für Gymnasien, Oberstufe und Aufgabensammlung Ernst Klett Verlag, 2005, 2010																
Aufgabentitel	Das Ökosystem See																
Aufgabenstamm und Impuls	<p>Durch menschliche Einflüsse (z.B. Dünger) können vermehrt Nährstoffe in Gewässer gelangen. Diese führen zu einer Veränderung des Sauerstoffgehalts im See und können eine erhöhte Produktion von Algen (Algenblüte) bedingen.</p> <p>Im Freilandpraktikum hat Paul den Nitratgehalt des Sengsees und von seinem Zufluss alle 30 Tage gemessen. An Tag 1 betrug der Nitratgehalt für den Zufluss 60 mg/l. Seine weiteren Messergebnisse für den See und den Zufluss hat er in die nebenstehende Tabelle eingetragen.</p>																
			<table border="1"> <thead> <tr> <th>Nitratgehalt in mg/l für den <u>Zufluss</u></th> <th>Nitratgehalt in mg/l für den <u>See</u></th> </tr> </thead> <tbody> <tr> <td>60</td> <td>30</td> </tr> <tr> <td>30</td> <td>10</td> </tr> <tr> <td>40</td> <td>5</td> </tr> <tr> <td>80</td> <td>30</td> </tr> <tr> <td>25</td> <td>50</td> </tr> <tr> <td>25</td> <td>40</td> </tr> </tbody> </table>	Nitratgehalt in mg/l für den <u>Zufluss</u>	Nitratgehalt in mg/l für den <u>See</u>	60	30	30	10	40	5	80	30	25	50	25	40
Nitratgehalt in mg/l für den <u>Zufluss</u>	Nitratgehalt in mg/l für den <u>See</u>																
60	30																
30	10																
40	5																
80	30																
25	50																
25	40																
Fragestellung	Wählen Sie eine sinnvolle Skalierung für den Nitratgehalt in mg/l auf der Ordinate und konstruieren Sie mithilfe der Tabellenwerte ein vollständig beschriftetes Diagramm, das die Nitratgehalte für den Zufluss und den See in Abhängigkeit vom Tag der Probenahme darstellt.																
Antwort																	

Item See09	Translation und Transformation (MERIII)	Ökologie	Diagramm
Quelle	NATURA Biologie für Gymnasien, Oberstufe Ernst Klett Verlag, 2010		
Aufgabentitel	Das Ökosystem See		
Aufgabenstamm und Impuls	s. Kapitel 4.2.2		
Fragestellung			
Antwort	 <p>Handwritten notes below the chart:</p> <p>März: 20 cm Mai: 2 m August: 4 m</p>		

Item Flo08	Informationsentnahmenahnahe Integration (MERI)	Ökologie	Diagramm
Quelle	LINDER Biologie, Gesamtband Schroedel, 2010, S. 401, Abb. 401.2: <i>Populationsentwicklung in einem Mehrartensystem vor, während und nach Einsetzen eines Fisches in die Biozönose</i>		
Aufgabentitel	Wasserflöhe in Teich und Labor		
Aufgabenstamm und Impuls	In einem See leben u.a. Fische, Zooplankton und verschiedene Algen. Das Zooplankton ernährt sich als Räuber hauptsächlich von kleinen Algen (Beute). In einem stabilen Ökosystem (Teich) ist die Räuber-Beute-Beziehung ausgeglichen. In Laborexperimenten können die Populationsdichten und somit die Anzahl der lebenden Organismen in einer Gruppe artgleicher Individuen, untersucht werden.		
	<p>Michl beobachtet vier verschiedene Zooplanktonarten in einem Becken. Nach ca. 35 Tagen setzt er einen Fisch ein. Er zählt alle vier Tage die Individuenzahl der Zooplanktonarten und erstellt das untenstehende Diagramm. Michl fragt sich, wie sich der Einsatz des Fisches auf die Individuenzahl auswirkt.</p>		
			
Fragestellung	Vergleichen Sie den Verlauf der vier Kurven über die gesamte Zeitspanne von 59 Tagen. Diskutieren Sie Michls Ergebnisse der Untersuchung und beantworten Sie Michls Frage.		
Antwort	<p><i>Die Arten Alonella und Ceriodaphnia liegen in den ersten 35 Tagen mehr oder weniger konstant vor. Danach steigt die Individuenzahl von Alonella von ca. 500 Individuen auf ein Maximum von ca. 700 Individuen an und sinkt ab Tag 51 wieder auf den Ausgangswert</i></p> <p><i>Nach Einsatz des Fisches steigt die Individuenzahl von Alonella stark an und sinkt nach Entfernen des Fisches ab Tag 51 wieder auf den Ausgangswert. Die Art Hytella wird durch den Fischeinsatz nicht beeinflusst.</i></p> <p><i>Die Individuenzahl von Ceriodaphnia geht gegen null, was darauf schließen lässt, dass Ceriodaphnia vom Fisch gefressen wird. Daraus lässt sich schlussfolgern, dass Ceriodaphnia Daphnia frisst, da mit Erholung des Fraßfeindes Ceriodaphnia die Daphnienzahl wieder abnimmt.</i></p>		

Kontext Stoffwechsel

Item Anp03	Informationsentnahmenahene Integration (MERI)	Stoffwechsel	Diagramm																														
Quelle	Abitur 2011 Biologie LK G9, 2005-32 STARK Verlag																																
Aufgabentitel	Anpassung an besondere Standorte																																
Aufgabenstamm und Impuls	Für die Fotosynthese benötigen Pflanzen Wasser und Kohlenstoffdioxid. Deshalb haben sie für die Anpassung an extreme Standorte verschiedene Überlebensstrategien entwickelt. Auf einer Tour durch den Botanischen Garten erfährt Lisa, dass in Florida die Orchidee <i>Catasetum</i> auf Baumkronen wächst. In Trockenzeiten wirft <i>Catasetum</i> ihre Blätter ab. Auf einer Informationstafel ist das untenstehende Diagramm abgebildet. Es zeigt die Kohlenstoffdioxid-Aufnahme bzw. -Abgabe von <i>Catasetum</i> mit und ohne Blätter in Abhängigkeit von der Uhrzeit.																																
	 <table border="1" data-bbox="587 922 1353 1330"> <caption>Data points from the graph</caption> <thead> <tr> <th>Uhrzeit (h)</th> <th>Catasetum mit Blätter (g/h)</th> <th>Catasetum ohne Blätter (g/h)</th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td></tr> <tr><td>3</td><td>-2</td><td>-2</td></tr> <tr><td>6</td><td>-1</td><td>-1</td></tr> <tr><td>9</td><td>0</td><td>-1</td></tr> <tr><td>12</td><td>5</td><td>-2</td></tr> <tr><td>15</td><td>20</td><td>-5</td></tr> <tr><td>18</td><td>10</td><td>-4</td></tr> <tr><td>21</td><td>5</td><td>-3</td></tr> <tr><td>24</td><td>0</td><td>0</td></tr> </tbody> </table>			Uhrzeit (h)	Catasetum mit Blätter (g/h)	Catasetum ohne Blätter (g/h)	0	0	0	3	-2	-2	6	-1	-1	9	0	-1	12	5	-2	15	20	-5	18	10	-4	21	5	-3	24	0	0
Uhrzeit (h)	Catasetum mit Blätter (g/h)	Catasetum ohne Blätter (g/h)																															
0	0	0																															
3	-2	-2																															
6	-1	-1																															
9	0	-1																															
12	5	-2																															
15	20	-5																															
18	10	-4																															
21	5	-3																															
24	0	0																															
Fragestellung	Vergleichen Sie für Lisa die beiden Kurvenverläufe. Begründen Sie die Aufnahme und Abgabe von Kohlenstoffdioxid bei der Orchideenart <i>Catasetum</i> mit und ohne Blätter.																																
Antwort	_____																																

Item Enz05	Konstruktionsnahe Integration (MERII)	Stoffwechsel	Schema
Quelle	Biologie Oberstufe, Gesamtband		
	Cornelsen, 2011, S. 68, <i>Abbildung 3: Ablauf einer enzymatisch katalysierten Reaktion, allgemeines Schema</i>		
Aufgabentitel	Enzyme		
Aufgabenstamm und Impuls	Forscher Arne erklärt seinen Studierenden den Ablauf einer durch Enzyme katalysierten Reaktion. Die Reaktion findet im aktiven Zentrum (taschenförmige Einbuchtung) statt. Enzym und Substrat passen wie Schlüssel und Schloss zusammen. Es bildet sich ein Enzym-Substrat-Komplex. Am Ende der Reaktion werden die Produkte freigesetzt und das Enzym bleibt unverändert zurück.		
	 <p>Das Diagramm zeigt den Ablauf einer enzymatisch katalysierten Reaktion in drei Schritten:</p> <ol style="list-style-type: none"> Ein Enzym (Schloss) und ein Substrat (Schlüssel) sind getrennt dargestellt. Ein leeres rechteckiges Feld befindet sich links neben dem Enzym, ein weiteres rechts neben dem Substrat. Das Substrat bindet an das aktive Zentrum des Enzyms, wodurch ein Enzym-Substrat-Komplex entsteht. Ein leeres rechteckiges Feld befindet sich unter dem Komplex. Die Reaktion ist abgeschlossen, die Produkte sind freigesetzt, und das Enzym bleibt unverändert zurück. Ein leeres rechteckiges Feld befindet sich rechts neben den Produkten, ein weiteres unter dem Enzym. 		
Fragestellung	Beschriften Sie das untenstehende Schema mithilfe von Arnes Erklärung und ergänzen Sie das fehlende Bild.		
Antwort	_____		

Item Blu05	Konstruktionsnahe Integration (MERII)	Stoffwechsel	Schema
------------	---------------------------------------	--------------	--------

Quelle

Unterricht Biologie, Friedrich Verlag, Heft Nr. 389 „Blut“, S. 39, *Abbildung: Das AB0-Blutgruppen-System*





Aufgabentitel

Blut

Aufgabenstamm und Impuls

Das Blut setzt sich aus den Blutzellen und flüssigem Blutplasma zusammen. Auf der Oberfläche der roten Blutkörperchen befinden sich besonders geformte Eiweißmoleküle, die Antigene, die eine Immunantwort auslösen. Diese können sich mit Antikörpern verbinden.





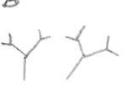

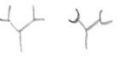
Resi geht Blutspenden. Aus dem Biologieunterricht weiß sie, dass sich Menschen in ihren Blutgruppen voneinander unterscheiden. Die Blutgruppen kommen durch Antigene auf den roten Blutkörperchen zustande. Erhält ein Empfänger eine nicht zu seiner Blutgruppe passende Blutinfusion, tritt eine lebensbedrohliche Verklumpung des Blutes ein. Die Verklumpung basiert auf einer Antigen-Antikörper-Reaktion, bei der sich Antikörper mit den dazu passenden Antigenen verbinden. Im Internet findet Resi das untenstehende Schema. Im Text liest sie dazu folgenden Satz: „Damit es zu keiner lebensbedrohlichen Verbindung der Antikörper mit den Antigenen kommen kann, passen diese bei den verschiedenen Blutgruppen entweder nicht zusammen oder es fehlen Antigene auf den roten Blutkörperchen bzw. die Antikörper im Serum.“

Antigene (Erythrozyten)	A 	B 		0 
Antikörper (Serum, der wässrige Bestandteil des Blutes)		A 	keine	
Blutgruppe	A		A/B	0

Fragestellung

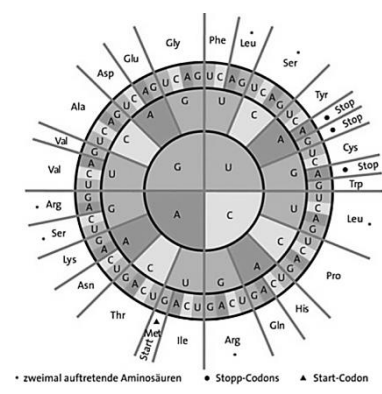
Vervollständigen Sie mithilfe des Textes das Schema.

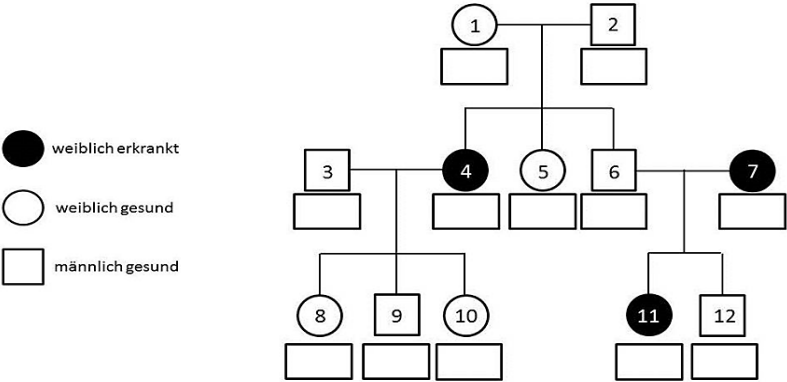
Antwort

Antigene (Erythrozyten)	A 	B 		0 
Antikörper (Serum, der wässrige Bestandteil des Blutes)	B 	A 	keine	A u. B 
Blutgruppe	A	B	A/B	0

Kontext Genetik

Item Dna02	Informationsentnahmenahene Integration (MERI)	Genetik	Schema
Quelle	Biologie Oberstufe, Gesamtband Cornelsen, 2011, S. 157, <i>Abbildung 2: Code-„Sonne“ – Schemadarstellung des genetischen Codes</i>		
Aufgabentitel	<p>Erbe, Umwelt, Merkmal</p> <p>Auf der RNA (Ribonucleinsäure) sind Baupläne zur Herstellung von Proteinen gespeichert. Bei der Translation der RNA werden Aminosäuren zu einem Protein verkettet.</p> <p>Forscherin Monika erklärt Tine, dass immer drei Basen auf der RNA eine Einheit (sogenannte Codons) bilden und für eine Aminosäure stehen. Die Translation beginnt mit dem Start-Codon AUG für die Aminosäure Methionin (Met). Die verschiedenen Codons und die zugehörige Aminosäure lassen sich mit der nebenstehende Code-Sonne darstellen (= genetischer Code). Die Codons werden immer von innen nach außen gelesen.</p>		
Fragestellung	Wandeln Sie mithilfe der Code-Sonne die folgende Kette von Aminosäuren in einen auf der RNA möglichen gespeicherten Code um. Met – Gly – Ala – Asn – Val – Val – Cys – Pro – Thr		
Antwort	<p>In Klammern sind immer die möglichen Aminosäuren eines Codons angegeben:</p> <p>AUG - GG(G/A/C/U) – GC(G/A/C/U) – AA(C/U) – GU(G/A/C/U) – GU(G/A/C/U) – UG(U/C) – CC(G/A/C/U) – AC(G/A/C/U)</p>		



Item Pku08	Konstruktionsnahe Integration (MERII)	Genetik	Schema
Quelle	Biologie Oberstufe, Gesamtband Cornelsen, 2011		
Aufgabentitel	Phenylketonurie (PKU)		
Aufgabenstamm und Impuls	Phenylketonurie ist eine Erbkrankheit, die zu Stoffwechselstörungen und schließlich zu schwerer geistiger Behinderung führt. PKU wird autosomal-rezessiv vererbt, d.h. kranke Personen zeigen den Genotyp aa, Gesunde den Genotyp AA. Bei Überträgern der Krankheit spricht man von heterozygot, sie zeigen den Genotyp Aa. Der untenstehende Stammbaum gehört zur Familie Huber und zeigt die Vererbung von PKU über die Generationen hinweg.		
			
Fragestellung	Ordnen Sie den Personen 1 bis 12 die Genotypen AA, Aa oder aa zu und tragen Sie diese in die dafür vorgesehenen Kästchen ein.		
Antwort	Einzutragen in den Stammbaum		

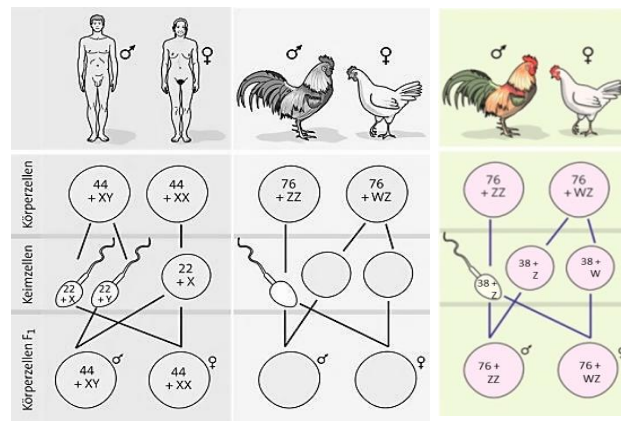
Item Erb05	Translation und Transformation (MERIII)	Genetik	Schema
------------	---	---------	--------

Quelle Biologie Oberstufe, Gesamtband
Cornelsen, 2011, S. 175, *Abbildung 2: Geschlechtschromosomen und Bestimmung des Geschlechts nach dem XY-Typ und dem WZ-Typ*

Aufgabentitel Erbe, Umwelt, Merkmal

Timo und Annabell befinden sich im Labor für Humangenetik. Forscher Alfons erklärt, dass die Geschlechtsbestimmung beim Menschen mithilfe der XY-Chromosomen erfolgt. Der Mann ist heterogametisch (XY) und die Frau homogametisch (XX). Anders als beim Menschen erfolgt die Geschlechtsbestimmung bei Hühnern über die WZ-Chromosomen. Das Weibchen ist heterogametisch (WZ) und das Männchen ist homogametisch (ZZ).

Das untenstehende Schema zeigt die Geschlechtschromosomen und die Bestimmung des Geschlechts bei Mensch und Huhn.



Fragestellung Ergänzen Sie die fehlende Anzahl und die Chromosomen in den Keimzellen und Körperzellen F₁ beim Huhn.

Antwort Siehe oben, nebenstehend

14.1.2 TESTHEFTDESIGN PILOTIERUNG

TH 1	TH 2	TH 3	TH 4	TH 5	TH 6	TH 7	TH 8	TH 9	TH 10
2	3	4	5	6	7	8	9	10	11
6	7	8	9	10	11	12	13	14	15
12	13	14	15	16	17	18	19	20	21
25	26	27	28	29	30	31	1	2	3
26	27	28	29	30	31	1	2	3	4
28	29	30	31	1	2	3	4	5	6

TH 11	TH 12	TH 13	TH 14	TH 15	TH 16	TH 17	TH 18	TH 19	TH 20
12	13	14	15	16	17	18	19	20	21
16	17	18	19	20	21	22	23	24	25
22	23	24	25	26	27	28	29	30	31
4	5	6	7	8	9	10	11	12	13
5	6	7	8	9	10	11	12	13	14
7	8	9	10	11	12	13	14	15	16

TH 21	TH 22	TH 23	TH 24	TH 25	TH 26
22	23	24	25	26	27
26	27	28	29	30	31
1	2	3	4	5	6
14	15	16	17	18	19
15	16	17	18	19	20
17	18	19	20	21	22

TH 27	TH 28	TH 29	TH 30	TH 31
28	29	30	31	1
1	2	3	4	5
7	8	9	10	11
20	21	22	23	24
21	22	23	24	25
23	24	25	26	27

14.1.3 TESTHEFTDESIGN HAUPTSTUDIE

Testheft 1					
A	Item	B	Item	C	Item
Öko I	See_02	SW I D	Atm_04	Gen II	PKU_08
Öko III	See_09	SW III D	Blu_15	Gen II	Epo_03
Gen II	Blu_05	SW II S	Enz_05	Öko II	See_07

Testheft 2					
A	Item	B	Item	C	Item
SW I D	Atm_04	Gen I	PKU_08	SW III S	Enz_02
SW III D	Blu_15	Gen III	Epo_03	SW I S	Anp_09
SW II S	Enz_05	Öko II	See_07	SW II D	Atm_03

Testheft 3					
A	Item	B	Item	C	Item
Gen II	PKU_08	SW III S	Enz_02	Gen III	PKU_07
Gen II	Epo_03	SW I S	Anp_09	Gen III	Blu_08
Öko II	See_07	SW II D	Atm_03	Öko I	Öko_02

Testheft 4					
A	Item	B	Item	C	Item
SW III S	Enz_02	Gen III	PKU_07	SW III S	Anp_07
SW I S	Anp_09	Gen III	Blu_08	SW II S	Enz_07
SW II D	Atm_03	Öko I	Öko_02	SW I D	Anp_03

Testheft 5					
A	Item	B	Item	C	Item
Gen III	PKU_07	SW III S	Anp_07	Öko III	Flo_09
Gen III	Blu_08	SW II S	Enz_07	Öko II	Flo_04
Öko I	Öko_02	SW I D	Anp_03	Gen I	Blu_04

Testheft 6					
A	Item	B	Item	C	Item
SW III S	Anp_07	Öko III	Flo_09	SW III D	Blu_11
SW II S	Enz_07	Öko II	Flo_04	SW II D	Epo_02
SW I D	Anp_03	Gen I	Blu_04	SW I S	Atm_11

Testheft 7					
A	Item	B	Item	C	Item
Öko III	Flo_09	SW III D	Blu_11	Öko II	Öko_04
Öko II	Flo_04	SW II D	Epo_02	Öko I	Flo_08
Gen I	Blu_04	SW I S	Atm_11	Gen III	Erb_05

Testheft 8					
A	Item	B	Item	C	Item
SW III D	Blu_11	Öko II	Öko_04	SW II D	Epo_05
SW II D	Epo_02	Öko I	Flo_08	SW I D	Anp_04
SW I S	Atm_11	Gen III	Erb_05	SW III S	Blu_02

Testheft 9					
A	Item	B	Item	C	Item
Öko II	Öko_04	SW II D	Epo_05	Gen I	Blu_09
Öko I	Flo_08	SW I D	Anp_04	Gen I	DNA_02
Gen III	Erb_05	SW III S	Blu_02	Öko III	Öko_06

Testheft 10					
A	Item	B	Item	C	Item
SW II D	Epo_05	Gen I	Blu_09	SW II S	Blu_06
SW I D	Anp_04	Gen I	DNA_02	SW I S	Enz_04
SW III S	Blu_02	Öko III	Öko_06	SW III D	Apr_01

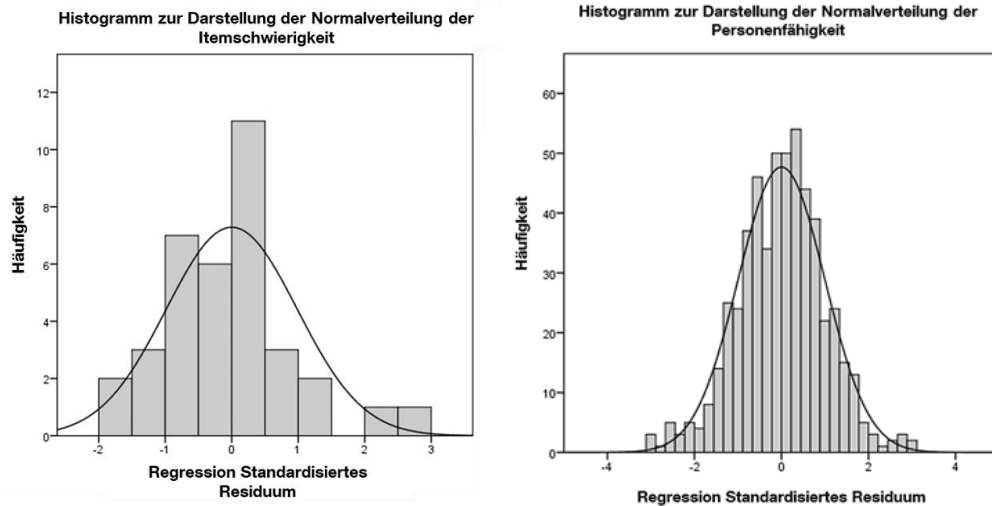
Testheft 11					
A	Item	B	Item	C	Item
Gen I	Blu_09	SW II S	Blu_06	Öko I	See_02
Gen I	DNA_02	SW I S	Enz_04	Öko III	See_09
Öko III	Öko_06	SW III D	Apr_01	Gen II	Blu_05

Testheft 12					
A	Item	B	Item	C	Item
SW II S	Blu_06	Öko I	See_02	SW I D	Atm_04
SW I S	Enz_04	Öko III	See_09	SW III D	Blu_15
SW III D	Apr_01	Gen II	Blu_05	SW II S	Enz_05

14.2 ERGÄNZUNGEN ZUR ERGEBNISDARSTELLUNG

Residuenverteilung

Die Normalverteilung der Residuen wird durch eine grafische Testung bestätigt.

**Kreuztabelle mit Kennwerten der Schwierigkeit für die Ausprägungen im Modell**

MER-Integration	Kontext	Repräsentationstyp	M	SE	N	
MERI	Ökologie	Diagramm	-,0730	,51966	5	
		Gesamt	-,0730	,51966	5	
	Stoffwechsel	Diagramm	,5633	1,20464	6	
		Schemazeichnung	,1423	,39028	6	
	Gesamt	Gesamt	,3528	,88158	12	
		Genetik	Schemazeichnung	-,9650	,44562	3
	Gesamt	Gesamt	Gesamt	-,9650	,44562	3
			Diagramm	,2741	,97161	11
		Stoffwechsel	Schemazeichnung	-,2268	,67186	9
			Gesamt	,0487	,86734	20
MERII	Ökologie	Diagramm	1,0445	1,10737	4	
		Gesamt	1,0445	1,10737	4	
	Stoffwechsel	Diagramm	1,0247	,59229	3	
		Schemazeichnung	,8190	,45627	3	
	Gesamt	Gesamt	,9218	,48609	6	
		Genetik	Schemazeichnung	,4977	,58936	6
	Gesamt	Gesamt	Gesamt	,4977	,58936	6
			Diagramm	1,0360	,85451	7
		Stoffwechsel	Schemazeichnung	,6048	,54309	9
			Gesamt	,7934	,70583	16
Gesamt	Ökologie	Diagramm	,4237	,97045	9	
		Gesamt	,4237	,97045	9	
	Stoffwechsel	Diagramm	,7171	1,02366	9	
		Schemazeichnung	,3679	,51158	9	
	Gesamt	Gesamt	,5425	,80533	18	
		Genetik	Schemazeichnung	,0101	,89531	9
	Gesamt	Gesamt	Gesamt	,0101	,89531	9
			Diagramm	,5704	,97933	18
		Stoffwechsel	Schemazeichnung	,1890	,73093	18
			Gesamt	,3797	,87335	36

Multiple Regression auf Itemparameter (M_IRT)

Komponenten	nicht standardisierte Koeffizienten		standardisierte Koeffizienten β	T	p
	B	SE			
Regressionskonstante B_0	0,189	0,204		0,928	0,360
Diagramm	0,381	0,288	0,221	1,324	0,194
Varianzaufklärung: $R^2 = 0,05$, $R^2_{korr} = 0,22$ ($p > 0,05$)					

Komponenten	nicht standardisierte Koeffizienten		standardisierte Koeffizienten β	T	p
	B	SE			
Regressionskonstante B_0	0,789	0,200		3,967	0,000
MER-Integration	-0,745	0,268	-0,430	-2,775	0,009
Varianzaufklärung: $R^2 = 0,19$, $R^2_{korr} = 0,16$ ($p < 0,01$)**					

Komponenten	nicht standardisierte Koeffizienten		standardisierte Koeffizienten β	T	p
	B	SE			
Regressionskonstante B_0	1,080	0,216		5,010	0,000
MERI-Integration	-0,917	0,257	-0,529	-3,567	0,001
Genetik	-0,764	0,295	-0,384	-2,592	0,015
Varianzaufklärung: $R^2 = 0,32$, $R^2_{korr} = 0,28$ ($p < 0,01$)**					

Ergebnisse der Tests auf Zwischensubjekteffekte in Bezug auf den Studiengang

MER-Integration	Testergebnis
MERI-Integration	$F(5,539) = 1,379, p = 0,23$
MERII-Integration	$F(5,539) = 2,322, p = 0,04^*$

Aufgabenkontext	Testergebnis
Ökologie	$F(4,505) = 1,208, p = 0,31$
Stoffwechsel	$F(4,505) = 2,983, p = 0,02^*$
Genetik	$F(4,505) = 2,677, p = 0,03^*$

Repräsentationstyp	Testergebnis
Diagramm	$F(5,541) = 1,673, p = 0,14$
Schemazeichnung	$F(5,541) = 1,579, p = 0,16$

Korrelationen mit den Schulnoten, getrennt nach MERI- und MERII-Integration

		Welche Zeugnis-Schulnote in Punkten hatten Sie in Biologie?	Welche Zeugnis-Schulnote in Punkten hatten Sie in Mathematik?	WLEs_MERI
Welche Zeugnis-Schulnote in Punkten hatten Sie in Biologie?	Korrelation nach Pearson	1	,142**	,108*
	Signifikanz (2-seitig)		,001	,013
	N	524	524	522
Welche Zeugnis-Schulnote in Punkten hatten Sie in Mathematik?	Korrelation nach Pearson	,142**	1	,026
	Signifikanz (2-seitig)	,001		,554
	N	524	538	536
WLEs_MERI	Korrelation nach Pearson	,108*	,026	1

		Welche Zeug- nis-Schulnote in Punkten hatten Sie in Biologie?	Welche Zeug- nis-Schulnote in Punkten hatten Sie in Mathematik?	WLEs_MERII
Welche Zeug- nis-Schulnote in Punkten hatten Sie in Biologie?	Korrelation nach Pearson	1	,142 ^{**}	,104 [*]
	Signifikanz (2-seitig)		,001	,017
	N	524	524	522
Welche Zeug- nis-Schulnote in Punkten hatten Sie in Mathematik?	Korrelation nach Pearson	,142 ^{**}	1	,038
	Signifikanz (2-seitig)	,001		,379
	N	524	538	536
WLEs_MERII	Korrelation nach Pearson	,104 [*]	,038	1

14.3 PUBLIKATIONEN

14.3.1 REFERIERTE BEITRÄGE ZU FORSCHUNGSTHEMEN

Beck, C. & Nerdel, C. (2016). Bildverständnis im Biologieunterricht diagnostizieren und fördern. *Unterricht Biologie*, 417(40), 38-43.

14.3.2 BEITRÄGE IN ABSTRACTBÄNDEN

Beck, C. & Nerdel, C. (2016). *Representational competence in integrating multiple external representations (MERs) in biology education*. In: Grace, M. et al. (eds.), 11th Conference of European Researchers in Didactics of Biology (ERIDOB). Karlstad: Karlstads Universitet.

Beck, C. & Nerdel, C. (2015). *Integration Multipler Externer Repräsentationen: Kompetenzmodellierung und der Einfluss von Aufgabenmerkmalen auf die Schwierigkeit*. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 95-96. Hamburg: Universität Hamburg.

Schwarz, S., Beck, C. & Nerdel, C. (2015). *Expertisedifferenziertes Bildverständnis bei der Integration von Multiplen Externen Repräsentationen*. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 117-118. Hamburg: Universität Hamburg.

Beck, C. & Nerdel, C. (2015). *Integration von Multiple External Representations in Biological Education*. In: Lampiselkä et al. (eds.), 11th Conference of the European Science Education Research Association (ESERA). Helsinki: University of Helsinki.

Beck, C. & Nerdel, C. (2015). *Raschanalyse von biologischen Aufgaben zum Umgang mit multiplen externen Repräsentationen (MER) zur Erfassung von schwierigkeitsgenerierenden Merkmalen im biologischen Kontext*. In: von Kotzebue, L., Dittmer, A., Möller, A., & Schmiemann, P. (Hrsg.), 17. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 26-27. München: Technische Universität München und LMU München.

Beck, C. & Nerdel, C. (2014). *Schwierigkeitsgenerierende Merkmale beim Umgang mit multiplen externen Repräsentationen (MER) im Biologieunterricht*. In: Chernyak, D., Möller, A., Dittmer, A., & Schmiemann, P. (Hrsg.), 16. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 42-43. Trier: Universität Trier.

Beck, C. & Nerdel, C. (2013). *Repräsentationen als schwierigkeitsgenerierende Merkmale bei Biologieaufgaben*. In: Mayer, J., Hammann, M., Wellnitz, N., Arnold, J., & Werner, M. (Hrsg.), 19. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 198-199. Kassel: Universität Kassel.

Beck, C. & Nerdel, C. (2013). *Repräsentationen als schwierigkeitsgenerierende Merkmale bei Aufgaben*. 40. Jahrestagung der Gesellschaft für Didaktik der Chemie und Physik, 82. München: LMU München und Technische Universität München.

Beck, C. & Nerdel, C. (2012). *Komplexität, Fachsprache, Aufgabenkontext - Schwierigkeitsgenerierende Merkmale bei Biologieaufgaben*. In: Jurgowiak, M., Zabel, J., Dittmer, A., Möller, A., & Schmiemann, P. (Hrsg.), 15. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 72-73. Leipzig: Universität Leipzig.

14.3.3 TAGUNGSPRÄSENTATIONEN

Beck, C. & Nerdel, C. (2017). *Fachsprache im naturwissenschaftlichen Unterricht: Multiple externe Repräsentationen (MER) im biologischen Kontext*. 5. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF). Universität Heidelberg, 13.03. – 15.03.2017.

Beck, C. & Nerdel, C. (2015). *Integration Multipler Externer Repräsentationen: Kompetenzmodellierung und der Einfluss von Aufgabenmerkmalen auf die Schwierigkeit*. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBIO, 95-96. Universität Hamburg, 14.09. – 17.09.2015.

Beck, C. & Nerdel, C. (2015). *Integration von Multiple External Representations in Biological Education*. 11th Conference of the European Science Education Research Association (ESERA). University of Helsinki, 31.08 – 04.09.2015.

Beck, C. & Nerdel, C. (2015). *Raschanalyse von biologischen Aufgaben zum Umgang mit multiplen externen Repräsentationen (MER) zur Erfassung von schwierigkeitsgenerierenden Merkmalen im biologischen Kontext*. 17. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie (FDdB) im VBIO. Technische Universität München und LMU München, 23.02. – 26.02.2015.

DANKSAGUNG

„Das Leben besteht in der Bewegung. Jede Bewegung verläuft in der Zeit und hat ein Ziel.“

-Aristoteles-

Ich möchte mich an dieser Stelle bei allen Menschen bedanken, die mich in den letzten Jahren unterstützt und auf unterschiedlichste Weise zum Gelingen dieser Arbeit beigetragen haben.

Zunächst danke ich Prof. Dr. Claudia Nerdel als meine wissenschaftliche Mentorin und Erstbetreuerin meiner Dissertation für ihren fachkundigen und immer vielperspektivischen Rat, ihre konstruktive Hilfestellung und den großen Freiraum, diese Arbeit gestalten zu können. Ihr entgegengebrachtes Vertrauen schenkte mir auch in schwierigen Phasen die nötige Kraft und Ausdauer diese Arbeit zum Abschluss zu bringen. Ich danke Prof. Dr. Julia Schwanewedel für die Zweitbetreuung dieser Arbeit.

Ein herzlicher Dank geht an meine wunderbaren Kolleginnen und Kollegen der FDLS-Arbeitsgruppe. Danke für die hervorragende Zusammenarbeit, die vielen geselligen Mittagessen und die angenehme Teamatmosphäre im und außerhalb des Büros. Danke, dass ihr auch in meiner Endphase immer Verständnis hattet und mir stets Mut zugesprochen habt. Ich danke den Kolleginnen und Kollegen der TP2-Arbeitsgruppe im Teach@TUM-Team. Danke für den kreativen, konstruktiven Austausch und dass ihr mir in den stressigen letzten Wochen stets den Rücken freigehalten habt. Im Besonderen danke ich Dr. Bettina Blasini für ihr warmherziges und immer offenes Ohr sowie Dr. Lena von Kotzebue, Dr. Patricia Schöppner und Michael Achter für die zusprechenden Worte und ihre Fähigkeit, mich stets zum Lachen zu bringen. Danke Dagmar Frick für den genialen, finalen Formatierungssupport.

Ich bedanke mich bei den wissenschaftlichen Hilfskräften Christine Sprater, Anna Schechert, Tobias Nöbauer und Christoph Weglöhner, die zuverlässig Daten erfasst und viele andere Aufgaben stets pflichtbewusst erledigt haben. Herzlichen Dank an alle Studierende, die an dieser Studie teilgenommen haben sowie den Ansprechpartnern der unterschiedlichen Universitäten und Fakultäten.

Mein tiefempfundener Dank gilt meiner Familie und meinen Freunden – Danke für eure Unterstützung auf allen Ebenen. Danke an meine Mädels, Jungs und Coaches von CrossFit Bewegt für die auspowernden WODs und Weightlifting-Abende. Meinem Ehemann Korbinian Schmidt danke ich für seine große emotionale Unterstützung, seine unendliche Geduld und Geborgenheit, aus der ich sehr viel Kraft für die Fertigstellung dieser Arbeit geschöpft habe.

Danke!

