Dissertation

Qualitative Eigenschaften von Multiple-Choice-Fragenitems und deren Zusammenhang mit empirischen Kennzahlen

Eine Analyse der Situation an der Medizinischen Universität Graz

eingereicht von

Dr. med. univ. Johannes Bernhardt-Melischnig

zur Erlangung des akademischen Grades

Doktor der Medizinischen Wissenschaft (Dr. scient. med.)

an der

Medizinischen Universität Graz

ausgeführt am

Institut für Medizinische Informatik, Statistik und Dokumentation

unter der Anleitung von

Univ.-Prof. Dr. Josef Smolle

Univ.-Prof.in Dr.in DIin Andrea Berghold

O. Univ.-Prof. Mag. Dr. Gilbert Reibnegger

2015

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig angefertigt und abgefasst, und jene Personen und Institutionen, die am Zustandekommen der Forschungsdaten beteiligt waren, namentlich genannt habe. Andere als die angegebenen Quellen habe ich nicht verwendet und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen habe ich als solche kenntlich gemacht. Die Arbeit an der Dissertation und daraus entstandener Publikationen wurde gemäß den Regeln der „Good Scientific Practice“ durchgeführt.

Dr. Johannes Bernhardt-Melischnig

Graz, am 3. September 2015

Vorwort

Ich möchte am Anfang dieser Arbeit die Gelegenheit nutzen den persönlichen Stellenwert, den diese Dissertation für mich besitzt, zu erläutern. Ich arbeite nun schon seit gut sieben Jahren an der Medizinischen Universität Graz und ebenso lange an Projekten, die mit der Lehre und der Evaluierung der Lehre zu tun haben. Seit einigen Jahren erweitert sich mein Arbeitsgebiet um das Thema hochwertiger Prüfungen in der medizinischen Ausbildung.

Das strukturierte und systematische Vorgehen bei der Betrachtung und Bewertung der Leistung der Studierenden, unabhängig für wen die Daten und Ergebnisse nützlich sind, ist eine sowohl interessante wie auch extrem wichtige Arbeit im Rahmen der Ausbildung. Wie auch in anderen Entwicklungsprozessen, müssen auch in der medizinischen Ausbildung unserer Studierenden die Ausgangslage sowie auch die Entwicklungen aufgezeichnet, ausgewertet und rückgemeldet werden, um den Prozess besser verstehen und optimieren zu können.

Da meine Formulierungen an dieser Stelle zu technisch klingen könnten, möchte ich als Gegenpol auf das Kapitel 2.3.4 mit interessanten Aspekten von Schuwirth und van der Vleuten verweisen. So vorteilhaft die Nutzung von Zahlen auch sein mag, wir müssen unseren *numerisch-quantitativen* Weg um andere Aspekte ergänzen und erweitern.

Die Beschäftigung mit diesen Themen ist spannend, vielfältig und teils sehr kontrovers, umso mehr sehe ich in dieser vorliegenden Arbeit meinen Einstieg in diesen interessanten Themenkomplex.

Danksagung

Ich möchte mich zu Beginn ganz herzlich bei meinem Dissertationskomitee, bestehend aus Prof. Dr. Josef Smolle, Frau Prof. Dr. Andrea Berghold und Prof. Dr. Gilbert Reibnegger bedanken. Diese drei Kolleginnen und Kollegen haben mir mit viel Geduld, Geschick und Professionalität den richtigen Weg gewiesen und mich in vielen Belangen direkt und indirekt unterstützt.

Ich möchte mich darüber hinaus natürlich auch ganz liebevoll bei meiner gesamten Familie, allen voran meiner Gattin Melanie und meiner Tochter Nina Anna-Maria bedanken, ohne deren Verständnis und Unterstützung diese Arbeit nie abzuschließen gewesen wäre.

In weiterer Folge möchte ich mich aber auch bei meiner Kollegin Frau Dr. Regina Riedl und meinem Vater Christian Bernhardt bedanken, die gerade in der Schlussphase meine Arbeit regelmäßig durchgesehen und gegengelesen haben.

Abschließend geht ein ebenso besonderer Dank an Herrn Hans-Christian Caluba, der für mich erster Ansprechpartner war und mir geduldig und gewissenhaft weitergeholfen hat, wenn ich Datensätze aus den unterschiedlichen an der Medizinischen Universität Graz genutzten Systemen gebraucht habe.

Ich möchte am Ende noch einmal betonen, dass diese Auflistung an Unterstützerinnen und Unterstützern nie vollständig sein kann und ich mich abschließend bei meinem kompletten privaten und beruflichen Umfeld mit allen Personen bedanken möchte.

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abkürzung	Erklärung
ECTS	„European Credit Transfer System“, ein europäisches System zur Übertragung und Akkumulierung von Studienleistungen[1]. Es erhöht die Transparenz und vereinfacht die Anerkennung von Leistungen im europäischen Hochschulraum.
IMS	„Item Management System“ oder in der Originalschreibweise „ItemManagementSystem“, ist ein System für das Item- und Klausurmanagement[2], welches von der heutigen UCAN-Organisation entwickelt wurde. Eine etwas detailliertere Beschreibung findet sich im Kapitel 4.6.5.
MC-Frage/ MC-Prüfung	“Multiple-Choice-Frage” oder „Multiple-Choice-Prüfung”, steht für ein Prüfungsformat mit geschlossenen Prüfungsfragen, bei dem mehrere vordefinierte Antwortoptionen zur Verfügung stehen. Details zu Definition und Aufbau finden sich im Kapitel 3.1.
OSCE/OSKE	„Objective structured clinical examination” oder „Objektiv strukturiertes klinisches Examen” steht für ein Prüfungsformat zur Überprüfung manueller/klinischer Fertigkeiten und wird im Kapitel 2.3.3 beschrieben.

[1] ECTS: http://ec.europa.eu/education/ects/ects_de.htm (abgerufen am 11. September 2015).
[2] IMS: https://www.ucan-assess.org/cms/de/tools/item-and-exam-management/ (abgerufen am 11. September 2015).

Abkürzung	Erklärung
PTM	„Progress Test Medizin“, ein interdisziplinärer Wissenstest für Studierende der Humanmedizin, von der Charité in Berlin erstellt und durchgeführt. Dieser Test wird im Kapitel 4.6.7 näher erklärt[3].
QMP	„Questionmark Perception“, ist eine Prüfungssoftware der Firma Questionmark, welche den kompletten Prüfungsprozess abbildet[4].

[3] PTM: http://ptm.charite.de/ (abgerufen am 8. Juli 2015).
[4] QMP: https://www.questionmark.com/de/content/questionmark-perception (abgerufen am 11. September 2015).

Zusammenfassung

In der vorliegenden Arbeit werden Multiple-Choice Fragenitems an der Medizinischen Universität Graz analysiert. Dazu wird ein großer Teil der in den Studienjahren 2011/2012 und 2013/2014 verwendeten Fragenitems im Fragenpool hinsichtlich unterschiedlicher Eigenschaften und deren Zusammenhänge untersucht. Die Unterschiede zwischen den einzelnen Abschnitten des Curriculums wurden besonders herausgearbeitet und abschließend mit den Ergebnissen des PTM verglichen.

Analysiert wurden alle Prüfungsdurchgänge von 13 Modulen, die Papier-basiert mit Multiple-Choice Fragen prüfen, und bei denen mehr als 25 Studierende bei jedem Durchgang angetreten sind. Aus 90 Durchgängen wurden so 4530 Fragenitems hinsichtlich mehrerer Eigenschaften und den Kennzahlen Item-Schwierigkeit und Item-Trennschärfe untersucht.

Die Ergebnisse lassen sich folgendermaßen zusammenfassen: Typ-K Fragen und Aussagen-basierte Fragen kommen vermehrt im Bereich der Vorklinik vor und sind meist schwieriger und trennschärfer. Im Gegenzug dazu kommen negative Formulierungen und Vignetten-Fragen vermehrt im klinischen Abschnitt vor und sind zudem meist leichter und weniger trennscharf. In Summe ergibt sich das Bild, dass es im klinischen Bereich schwerer ist, fokussierte und trennscharfe Fragenitems mit passendem Schwierigkeitsgrad zu erstellen, die dazu entsprechend gute Distraktoren haben. Dieses Bild deckt sich abschließend mit dem PTM-Ergebnis, bei dem Studierende der Medizinischen Universität Graz im klinischen Bereich unter dem Durchschnitt abschneiden.

Die Schulung und intensive Betreuung der Lehrenden in Bezug auf die Multiple-Choice-Fragen-Erstellung ist ebenso – mit einem klinischen Fokus – fortzusetzen, wie alternative Prüfungsformate diskutiert und evaluiert werden müssen. Zusätzlich müssen die in dieser – zumeist explorativen – Arbeit dargestellten Ergebnisse mit detaillierten Untersuchungen verifiziert und ergänzt werden.

Abstract

This dissertation focuses on the analysis of multiple-choice items at the Medical University of Graz. Item characteristics and their correlations were analyzed. Differences between three phases of the curriculum were of special interest, at the end, these differences were compared to the PTM results.

The analysis included all written exams of 13 study modules with more than 25 students taking part. Finally 90 exams with 4530 question items were analyzed, all with regard to item characteristics and item difficulty and discrimination index.

As a main result one can state, that questions of the complex type and statement type are more difficult and better discriminating and are more widely used in the pre-clinical phase. In contrast negative questions and question items with case vignettes are less difficult and discriminating and more widely used in the clinical phase of the curriculum. Concluding the results, you can summarize, that it is more difficult to write good, focused and discriminating questions, of appropriate difficulty and with precise distractors in the clinical phase. This picture perfectly meets and partly explains the PTM results, where students of the Medical University of Graz perform lower than average.

Intensive instruction and training in creating excellent multiple-choice items, especially for the clinical phase, is necessary, as well as discussing and evaluating other examination formats. Additionally all the results of the dissertation need verification and further more detailed analysis.

1. Einführung

1.1. Motivation und Fragestellungen

Im Zentrum dieser Dissertation steht eine explorative Untersuchung, die sich mit der speziellen Situation an der Medizinischen Universität Graz beschäftigt – es soll eine repräsentative Anzahl an Multiple-Choice-Fragen (kurz: MC-Fragen) für Prüfungen aus dem Humanmedizin-Curriculum analysiert werden. Im Detail geht es darum, die Verteilung bestimmter Eigenschaften und die Qualität der Fragenitems und der Antwortoptionen zu untersuchen. Dabei soll auch der Zusammenhang der Eigenschaften mit den berechneten Qualitätskennwerten der Fragenitems und Antwortoptionen untersucht werden.

Die Eigenschaften müssen dabei einfach und eindeutig zu bestimmen sein. Deshalb wurde auf das Einbeziehen von Eigenschaften verzichtet, welche sehr vom Frageninhalt und der Formulierung abhängig sind und zudem nur subjektiv zu bestimmen wären. Gerade das Vorhandensein von ungewollten Lösungshinweisen scheint leicht bewertbar zu sein, was jedoch täuscht.

Zum Hintergrund ist zu sagen, dass das MC-Format beliebt ist und häufig verwendet wird. Zudem gibt es aktuell an der Medizinischen Universität Graz zahlreiche Maßnahmen um die Qualität der MC-Prüfungen zu heben: Bereitstellung von Schulungsunterlagen, Anbieten von Schulungen im Rahmen der Internen Weiterbildung, individuelle Betreuung beim Erstellen von Prüfungsfragen und das Etablieren eines verpflichtenden Peer-Review-Prozesses. Auch die vorliegende Arbeit könnte als Teil dieser Qualitätsoffensive verstanden werden. Erstens wird die aktuelle Lage beschrieben, um eine valide Datenbasis für zukünftige Vergleiche zu haben.

Zweitens ist die Kenntnis der Charakteristika des aktuellen Fragenpools hilfreich für Schulungsmaßnahmen. Die Kernfrage ist schließlich, wie Eigenschaften und Kennwerte zusammenhängen, auch in Abhängigkeit des Studienjahres und des Studienabschnittes.

Die abschließende, praktische Fragestellung gilt vorzugsweise dem klinischen Abschnitt und den dort vorherrschenden Anwendungsfragen: Sind Anwendungsszenarien weniger eindeutig und damit schwieriger zu beschreiben als Sachverhalte im vorklinischen Bereich und haben sie am Ende einen niedrigeren Schwierigkeitsgrad als vorklinische Fragen? In diesem Zusammenhang werden die PTM-Ergebnisse, also die Ergebnisse des Progress Test Medizin der Berliner Charité, als Referenz herangezogen. Wie die Prüfungsfragen an der Medizinischen Universität Graz zu charakterisieren sind und wie sich die Eigenschaften auf die Abschnitte verteilen ergibt ein Bild innerhalb der Universität, der PTM bietet sich aber darüber hinaus als vergleichende Messung an.

Die Fragestellungen noch einmal detailliert zusammengestellt:

1. Wie lassen sich die MC-Fragen aus dem Pool der Medizinischen Universität Graz anhand der untersuchten und berechneten Eigenschaften charakterisieren?
2. In welchem Zusammenhang stehen die zuvor genannten Eigenschaften? Welche Item-Eigenschaften haben einen Einfluss auf die Kennwerte Item-Schwierigkeit und Item-Trennschärfe und wie groß ist dieser?
3. Gibt es bezüglich der ersten beiden Fragestellungen Unterschiede in Abhängigkeit der untersuchten Studienjahre oder der Studienabschnitte?
4. Haben der klinische Studienabschnitt und die dort vorherrschenden Fragen besondere Eigenschaften, die sich von den ersten beiden Studienabschnitten unterscheiden und damit ein Erklärungsmodell für die Ergebnisse des PTM-Tests liefern können?

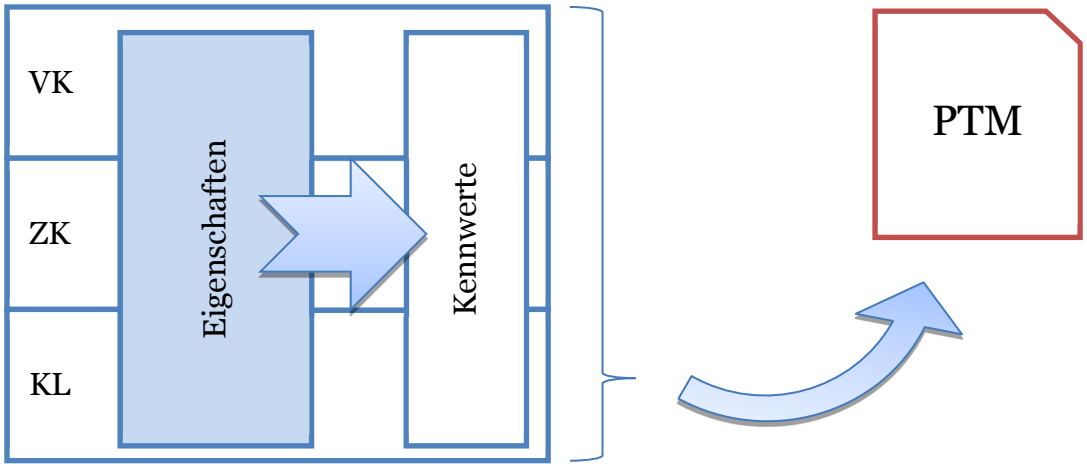

Abbildung 1: Grafische Übersicht über den groben Zusammenhang der Daten, die in dieser Arbeit verwendet werden: Item-Eigenschaften und Item-Kennwerte über die Abschnitte Vorklinik (VK), Zwischenklinik (ZK) und Klinik (KL) verteilt und den Zusammenhang der Ergebnisse mit dem PTM.

1.2. STRUKTUR DER DISSERTATION

Nach dem aktuell vorliegenden *Kapitel 1 – Einführung* schließt das *Kapitel 2 – Prüfen in der medizinischen Ausbildung* an. In diesem werden, nach einem kurzen Abschnitt zur medizinischen Ausbildung, die wichtigsten Eckpunkte des damit verbundenen Themas Prüfen erläutert. Die Ziele der Leistungsbeurteilung sind dabei ebenso enthalten wie die Prüfungsinhalte und -methoden.

Im anschließenden *Kapitel 3 – Prüfungsformat Multiple-Choice*, wird eben dieses Format erklärt: Wie es definiert ist, welchen Aufbau und welche Vor- und Nachteile es besitzt, welche Arten von MC-Fragen es gibt und wie anwendungsorientierte Fragen aussehen.

Das *Kapitel 4 – Erstellung und Verwendung von Multiple-Choice-Fragen* listet zahlreiche Richtlinien und Empfehlungen auf und diskutiert diese. Nach einem kurzen Abschnitt zum Thema Prüfungszusammenstellung widmet sich das Kapitel auch den Fragen der Qualitätssicherung. Neben dem Schulungsthema wird vor allem das Peer-Review-System erläutert.

Am Ende dieses Kapitels werden die statistischen Kennzahlen behandelt, um die Prüfungs-, Item-, oder Distraktorebene darzustellen.

Im *Kapitel 5 – Material und Methoden w*ird kurz auf die spezifische Situation an der Medizinischen Universität Graz eingegangen, um darauf folgend die Datenauswahl und -bearbeitung erklären zu können. Am Ende werden die verwendeten Methoden beschrieben.

Im *Kapitel 6 – Ergebnisse* werden die Ergebnisse tabellarisch und grafisch dargestellt und beschrieben, welche im *Kapitel 7 – Diskussion* anschließend zusammengefasst und der aktuellen Literatur gegenübergestellt werden. Eine Beschreibung der Grenzen dieser Arbeit sowie ein Ausblick schließen dieses Kapitel ab.

2. Prüfen in der medizinischen Ausbildung

2.1. Die Medizinische Ausbildung

Was macht einen guten Mediziner oder eine gute Medizinerin aus? Eine umfassende naturwissenschaftliche Basis sowie das Verstehen und Anwenden der grundlegenden Prinzipien der Natur sollten am Anfang der Ausbildung stehen. Ausreichende Kenntnisse über den menschlichen Körper und seiner funktionellen Zusammenhänge, Kenntnisse über diagnostische und therapeutische Möglichkeiten, ob technisches Instrumentarium oder pharmazeutische Präparate, ergänzen und erweitern diese Basis im weiteren Verlauf. Zu Beginn steht somit ein eher theoretisches Fundament, das nach und nach um klinisch-praktische Einheiten ergänzt wird. Doch um dem Ausbildungsziel näher zu kommen, sind auch andere Fähigkeiten unumgänglich notwendig: Der soziale Umgang mit den Mitmenschen, einfühlsames Handeln, kommunikative Fähigkeiten, aber auch professionelles Verhalten und selbstreflektiertes Handeln sind dabei nur einige der wichtigsten.

Es gibt zahlreiche Varianten, das Ziel der medizinischen Ausbildung zu beschreiben und zu definieren. Epstein und Hundert haben die ärztliche Kompetenz folgendermaßen beschrieben: „*Gewohnheitsmäßige, umsichtige Anwendung von Kommunikation, Wissen, technischen Fertigkeiten, klinischer Argumentation, Emotionen, Werten und Reflektion bei der alltäglichen Arbeit zum Wohl des Einzelnen als auch einer ganzen Bevölkerungsgruppe*“ (Epstein & Hundert 2002).

Anstatt nun aktuelle Gesetze[5] oder Ausbildungsziele[6] der medizinischen Universitäten aufzulisten, möchte ich als Einstieg auf die lebhafte Diskussion rund um die professionellen Kompetenzen verweisen. Epstein, bereits zuvor zitiert, hat zahlreiche Dimensionen aufgelistet. Neben klassischen Punkten wie Wissen, Anwendung von Wissen und Fertigkeiten finden sich auch nicht fachspezifische Aspekte, wie: Informationsmanagement, Erfahrung, Problemlösen, Umgang mit Unsicherheit, Zeitmanagement, bis hin zu Teamfähigkeit und emotionaler Intelligenz (Epstein & Hundert 2002). Diese Beschreibungen sind vergleichbar mit denen des ACGME Outcome Projektes (Swing 2007).

Simple Unterteilungen in Wissen, Fertigkeiten und Haltungen sind nur mehr bedingt hilfreich, die Beschreibung der ärztlichen Kompetenzen ist vieldimensional und komplex und vor allem auch nicht statisch (Huddle & Heudebert 2007). Sicher ist, dass Faktenwissen allein nicht ausreichen wird (Weih & Abrahamson 2006). Situatives Lernen, sozio-kulturell, aktiv-partizipierend, in einem klinischen Kontext, zusammen mit Patientinnen und Patienten ist entscheidend (Mann 2011; Jones et al. 2001).

Wie von Harden beschrieben (Harden 2000), gab es in den letzten Jahrzehnten zahlreiche neue *Bewegungen* in der medizinischen Ausbildung: Studierendenzentriertes und Problem-basiertes Lernen, sowie integrative Curricula. Viele dieser Neuerungen haben in den beginnenden 70er-Jahren mit vereinzelten Umsetzungen Einzug in die Ausbildung erhalten. Alle der drei oben genannten Bewegungen haben sich bis heute großteils durchgesetzt und sind nun allseits akzeptiert und anerkannt. Parallel dazu haben sich die Lerntechnologien von Folien und Dia-Projektionen, über einzeln eingesetzte Computer, bis hin zum web-basierten und mobilen Lernen weiterentwickelt.

[5] Ärzteausbildungsordnung 2015: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20009186 (abgefragt am 12. August 2015, 14:00 Uhr).
[6] Studienplan Version 13 mit den generellen Ausbildungszielen, sowie dem Qualifikationsprofil der Absolventinnen/Absolventen http://www.medunigraz.at/fileadmin/studieren/humanmedizin/pdf/studienplan_v13_01102014.pdf (abgefragt am 12. August 2015).

E-learning oder virtuelles Lernen genannt, zeigt am Beginn des dritten Jahrtausends zahlreiche Ausprägungen: Animationen, Simulationen, virtuelle Patienten, *gestreamte*[7] Lehrveranstaltungen, kollaboratives Arbeiten und Lernen, computerbasiertes Prüfen und ähnliches. Daneben darf jedoch nicht unerwähnt bleiben, dass sich das gesamte aktuelle Wissen unserer Zeit in eine Art *virtuelle Parallelwelt* verlagert. Online verfügbare Fachzeitschriften und Weiterbildungsangebote ersetzen damit die Notwendigkeit einer Präsenz in Bibliotheken und Schulungsräumen.

Aber auch beim Überprüfen der oben aufgezählten Dimensionen und Fähigkeiten gibt es zahlreiche Neuerungen: Es werden mehrere Prüfungsmethoden und -formate in Kombination eingesetzt, ebenso standardisierte und normale Patienten[8], daneben wird der Umgang mit medizinischer Literatur erfasst, ebenso die Teamfähigkeit oder die Entwicklung über die Zeit mit Hilfe von Portfolios (Epstein & Hundert 2002).

Hier sieht man bereits erste Schwierigkeiten, die sich anbahnen. Auf der einen Seite betrachten und beurteilen wir während der Ausbildung oft Fähigkeiten, die man leicht und komfortabel messen kann. Das Abrufen von Faktenwissen oder das Lösen von Rechenaufgaben kann relativ leicht mit einem schriftlichen Test überprüft werden, korrektes soziales und professionelles Verhalten, ein für den künftigen Beruf passendes Wertebild oder das Vorhandensein selbstreflektierten Verhaltens jedoch nur schwer. Das soll auf keinen Fall bedeuten, dass es sich Bildungseinrichtungen zu leicht machen, es soll nur festgehalten werden, dass man viele Fähigkeiten nur schwer bis gar nicht objektiv und zuverlässig messen kann.

[7] *streaming* bedeutet die gleichzeitige Übertragung und Wiedergabe von Audio- und Videodaten über ein Netzwerk. https://de.wikipedia.org/wiki/Streaming_Media (abgerufen am 12. August 2015).

[8] Unter *standardisierten Patienten* versteht man Personen, die einem Schauspieler ähnlich, eine Rolle gleichbleibend und standardisiert spielen. Der Ausdruck *normale Patienten* bedeutet dabei, dass es sich um wirkliche Patienten handelt, die ihr Einverständnis geben, eine Rolle in der medizinischen Lehre zu haben.

Wir müssen uns von der Kombination von Konstrukten, die einfach, stabil und unabhängig sein sollen, entfernen, ebenso von unserer Vorstellung, die Gesamtheit mit einer einzelnen Methode erfassen und beurteilen zu können. Abschließend – als Überleitung zum Prüfungsabschnitt und wie in 2.3.4 näher erläutert – ist Schuwirths Aussage angebracht, dass wir unsere psychometrischen Methoden erweitern sollen, wir müssen andere Wege finden, unsere Prüfungsergebnisse zu verteidigen (Schuwirth & van der Vleuten 2006).

2.2. GRÜNDE – ODER WARUM WIR PRÜFEN

Von Epstein wurden 2007 drei Assessment-Ziele in der medizinischen Ausbildung beschrieben: die Fähigkeiten der Lernenden durch Motivation und Hilfestellung zu optimieren, die Gemeinschaft vor inkompetenten Ärzten zu bewahren und das Bilden einer Basis für die Entscheidung über weiterführende Ausbildungen (Epstein 2007). Anhand dieser drei Ziele kann man auch gut die Einteilung in formative und summative Prüfungen sehen. Wenn Prüfungsergebnisse lediglich zur Orientierung über den Zwischenstand im Lernverlauf dienen und für die Teilnehmenden keine zwingenden Konsequenzen haben, spricht man von einer formativen Prüfung. Das würde Epsteins Hilfestellung entsprechen. Wenn es sich um sanktionierende Beurteilungen handelt, die oft am Ende eines Ausbildungsabschnittes oder der Ausbildung platziert sind und einschneidende Konsequenzen mit sich bringen, spricht man von summativen Prüfungen. Epstein macht den Abschluss einer Ausbildung und den Einstieg in eine weiterführende Ausbildung von dieser Prüfungsform abhängig. Der Erziehungswissenschafter Robert E. Stake[9] erklärt den Unterschied zwischen den verschiedenen Prüfungsformen sehr eindrucksvoll (Scriven 1991):

> *„When the cook tastes the soup, that's formative. When the guests tastes the soup, that's summative. “*

[9] Director of the „Center for Instructional Research and Curriculum Evaluation“, University of Illinois.

Eine etwas andere Einteilung wurde 2002 veröffentlicht, dabei wurden die Ziele für alle beteiligten Stakeholder getrennt aufgelistet (Epstein & Hundert 2002): Für die Studierenden, für die Institution oder das Curriculum und für die Gesellschaft. Ein interessanter Gedanke ist hierbei der, dass ein schlechtes Abschneiden bei der Überprüfung der Leistung nicht nur Rückschlüsse auf den Studierenden zulässt, sondern auch auf die Institution und die Seite der Lehrenden.

„*Assessment drives learning*" ist ein sehr oft gelesener und zitierter Satz: McLachlan hat dieses Statement genauer beleuchtet (McLachlan 2006). Van der Vleuten hat vier Arten beschrieben, wie diese Beeinflussung stattfindet: Durch den Inhalt, das Format, das Timing und durch das Feedback an die Kandidatinnen und Kandidaten (Van der Vleuten 1996). Auf der anderen Seite übt Wass im Jahr 2001 auch Kritik an diesem berühmten Statement (Wass et al. 2001). In Wirklichkeit sind Studierende *overloaded* und reagieren darauf mit einem Lernverhalten, nur diese Teile intensiv zu lernen, welche auch getestet werden. Um Lernen zu fördern, sollen Prüfungen formativ sein und den Studierenden Feedback geben. Pragmatisch gesehen sind Prüfungen eine Art Zaumzeug des Curriculums.

Ein wenig anders hat dies Krebs in seinen Publikationen beschrieben: „*Ergebnisse in Leistungsprüfungen sind nur von Nutzen, wenn daraus gültig (valide) und zuverlässig (reliabel) auf eine Leistung geschlossen werden kann, die über das Lösen der konkreten Prüfungsfragen/-aufgaben hinausreicht*", oder „*Beurteilung von Wissen, Fertigkeiten und Kompetenzen, welche bei der Lehrplan-Erstellung als Lernziele richtungsweisend für Lehrende und Studierende definiert worden sind!*" (Krebs 2008).

2.3. METHODEN – ODER WIE WIR PRÜFEN

Wenn wir uns mit der Frage beschäftigen wollen, wie wir die Leistung der Studierenden erfassen und beurteilen können, müssen wir die Psychologie heranziehen. Das Schlüsselwort ist hier die *Psychometrie* und bezeichnet das Messen von psychologischen Merkmalen.

Dabei spielen Testtheorien eine entscheidende Rolle, weil sie den Zusammenhang zwischen der Messung und dem eigentlichen Merkmal herstellen, ohne den man nicht von den Testergebnissen auf das wirkliche Vorliegen des Merkmals schließen könnte. Unser modellhafter Zusammenhang wird üblicherweise in der *klassischen Testtheorie* beschrieben.

In der psychologischen Diagnostik spielt der *psychologische Test* eine große Rolle, mit dem wir die Merkmale erfassen, die anschließend interpretiert werden sollen. Es gibt verschiedene Tests; hier handelt es sich um Leistungstests, begründet durch die Tatsache, dass es richtige und falsche Antworten gibt (beispielsweise im Vergleich zu Persönlichkeitstests) (Lienert & Raatz 1998).

2.3.1. GÜTEKRITERIEN

Da es sich hier um ein großes und umfangreiches Forschungsfeld handelt, möchte ich für meine Arbeit nur zwei Aspekte herausgreifen und näher erläutern. Wie jede wissenschaftliche Messmethode muss auch der psychologische Test gewisse Qualitätsmerkmale aufweisen – hier spricht man von Gütekriterien diagnostischer Testverfahren. Die wichtigsten Kriterien sind Objektivität, Reliabilität und Validität. Daneben gibt es noch einige weitere Nebenkriterien, wie: Fairness, Ökonomie, Transparenz, Unverfälschbarkeit, Zumutbarkeit und Normierung.

Die ersten beiden Kriterien möchte ich hier an dieser Stelle erläutern, die Validität wird in Kapitel 2.4 behandelt. *Objektivität* bedeutet, dass das Messergebnis unabhängig von der untersuchenden Person und ohne verzerrenden Einfluss der Rahmenbedingungen und der Untersuchungssituation zustande gekommen ist.

Reliabilität bedeutet Reproduzierbarkeit des Ergebnisses oder anders ausgedrückt, dass es sich um eine zuverlässige Messung handelt. Beim Wiederholen einer Messung soll ein gleiches oder zumindest sehr ähnliches Ergebnis auftreten, dieselben Rahmenbedingungen vorausgesetzt. Es gibt verschiedene Vorgehensweisen, die Reliabilität zu erheben: Man kann einen Test parallel durchführen (Paralleltest), einen Test in zwei Hälften teilen (Split-Half) oder ihn wiederholen (Retest) und anschließend beide Teile auf Ähnlichkeit untersuchen – allerdings sind alle Varianten wenig praxistauglich.

Daher behilft man sich mit der internen Konsistenz. Jedes Fragenitem wird sozusagen zum *Paralleltest* der übrigen Items. Bei dichotomen Items berechnet man die interne Konsistenz mit der Kuder-Richardson-Formel (Kuder & Richardson 1937), bei Items mit Intervallskalierung errechnet man die Kennzahl Cronbachs Alpha (Cronbach 1951).

Krebs von der Universität Bern spricht von der Beeinträchtigung der Messung durch *sachfremde Faktoren und Zufälligkeiten* (Krebs 2004). An zahlreichen Stellen existieren Auflistungen von Faktoren, die die Zuverlässigkeit positiv beeinflussen. An dieser Stelle soll Downing zitiert werden (Downing 2004):

- Große Anzahl an Fragenitems im Test
- Eindeutige Fragenformulierung
- Effektive Peer-Review-Runden bei der Itemerstellung
- Items von mittlerem Schwierigkeitsgrad
- Items vor der eigentlichen Verwendung testen

2.3.2. STANDARDSETZUNG

Neben dem Begriff der Standardsetzung, angelehnt an den englischen Begriff *standard setting*, spricht man auch vom Festlegen einer Bestehensgrenze. In unserem Verständnis, wie wir (kognitive) Leistung feststellen und beurteilen, spielt die Reduktion auf numerische Werte eine große Rolle. Vergleichbar mit der Normierung anderer psychologischer Tests, kann nur dann auf die realen Merkmale geschlossen werden, wenn eine Art Referenz oder Vergleichsmaßstab vorhanden ist (Wass et al. 2001). Diese Bestehensgrenze muss zuvor erarbeitet werden, um anschließend ein Studierenden-Testergebnis entsprechend beurteilen zu können. Liegt beispielsweise die Punktezahl über der Bestehensgrenze hat man bestanden, liegt sie darunter, hat man nicht bestanden.

Es gibt nun grundlegend zwei Arten, diese Grenze oder dieses Bezugssystem festzulegen, die normorientierte und die kriteriumsorientierte Variante. Erstere ist relativ und wird durch die Anzahl oder den Prozentsatz von Studierenden angegeben, die bestehen sollen. Dieser wird verwendet, wenn beispielsweise eine bestimmte Anzahl von Bewerberinnen oder Bewerber aufgenommen werden soll.

Bei der kriteriumsorientierten Variante wird die Anzahl oder der Prozentsatz an richtig gelösten Aufgaben oder richtig beantworteten Fragenitems festgelegt, um zu bestehen.

Bei der zweiten – absoluten – Methode ist es notwendig, dass Expertinnen und Experten vor jedem Testdurchgang die Mindestanforderungen festlegen. Da die mittlere Testschwierigkeit von Durchgang zu Durchgang variiert, ist es notwendig diese Standardfindung immer wieder von neuem durchzuführen, auch wenn es sich um einen zeitaufwendigen Prozess handelt. Es gibt zahlreiche Methoden. Der Fixe Prozentsatz, die Angoff- und die Hofstee-Methode sind nur drei, die häufiger verwendet werden und hier beispielhaft genannt werden sollen (Norcini 2003).

2.3.3. PRÜFUNGSFORMATE

Es gibt unbestritten sehr viele Prüfungsformate, manche von ihnen sind vielen geläufig; vor allem, wenn man auf die Einteilung schriftlich, mündlich und beobachtend oder praktisch verweist (Wass et al. 2001; Norcini & McKinley 2007). Andere wiederum, wie beispielsweise Portfolios, sind komplexer und aufwändiger, manche auch kaum mit der klassischen Testtheorie in Einklang zu bringen (L. Schuwirth & Van der Vleuten 2004).

Unter dem Begriff *Schriftlich* können viele unterschiedliche Formate subsumiert werden, welche sich durch schriftliches Eingeben oder Auswählen von Antworten auf Fragen oder Aufgaben kennzeichnen. Dabei ist es egal, ob dies auf Papier (Paper-Pencil und im Folgenden als *Papier-basiert* beschrieben) oder mittels Computer (oder im Folgenden als Computer-basiert beschrieben) geschieht. Schuwirth und Van der Vleuten (L. W. T. Schuwirth & Van der Vleuten 2004) haben einen Raster geschaffen, in dem sie *stimulus* und *response* gegenüberstellen. Beim Stimulus-Format unterscheiden sie zwischen reinen Fragen zu Faktenwissen und angewandten-klinischen Fragen, hingegen beim Response-Format zwischen *geschlossenen* MC-Fragen und *offenen* Fragenformaten, wie beispielsweise Short- oder Long-Essay Formaten.

Daneben gibt es noch zahlreiche andere Prüfungsformate mit ihren spezifischen Eigenschaften sowie Vor- und Nachteilen. Mündliche Prüfungsgespräche bieten sich an, wenn sprachlich-kommunikative und argumentative Fähigkeiten im Vordergrund stehen. In diesem Format ist ein flexibles Reagieren auf den Prüfungsverlauf möglich. Auf ein strukturiertes Vorgehen, um die Objektivität zu gewährleisten, muss jedoch Wert gelegt werden. Daneben fokussiert sich das relativ neue Prüfungsformat *Objektiv strukturiertes klinisches Examen* – kurz OSKE – (Davis 2003; Newble 2004) auf die Überprüfung praktischer Fertigkeiten in einem künstlichen Setting. Eine OSKE-Station ist einer umschriebenen Aufgabe gewidmet und wird dort beobachtet und – meist mittels einer Checkliste – beurteilt.

Üblicherweise werden dabei zahlreiche OSKE-Stationen an einem Prüfungstag mit Hilfe eines Rotationsschemas geprüft. Gelegentlich kann es notwendig sein, dass Patienten gespielt werden müssen. Dies geschieht mittels *Standardisierter Patientinnen und Patienten*, die versuchen, das Verhalten echter Patientinnen und Patienten vollständig, realistisch und zuverlässig zu verkörpern, was mitunter unterschiedlich herausfordernd sein kann. Dieses Format hat den Vorteil, „clinical performance", d. h. Fertigkeiten und Handlungskompetenzen – je nach Anzahl der Stationen auch zuverlässig – prüfen zu können, nachteilig wirken sich jedoch die benötigten Zeit- und Personalressourcen aus.

Möchte man hingegen Fertigkeiten und Fähigkeiten in einem natürlichen und realistischen Umfeld beobachten und beurteilen, hat man mehrere Möglichkeiten von strukturierten Beobachtungen. Hier wären beispielsweise die Formate Mini-CEX (Mini Clinical Evaluation Exercise) und DOPS (Direct Observation of Procedural Skills) zu nennen, wobei anzumerken ist, dass diese für die ärztliche Weiterbildung konzipiert wurden (Berendonk & Beyeler 2004). Das Mini-CEX-Format wurde vom American Board of Internal Medicine (Norcini et al. 1995), DOPS vom Royal College of Physicians (Wragg et al. 2003) eingeführt. Sie stellen sozusagen Momentaufnahmen realer Arzt-Patienten-Interaktionen dar und zeichnen sich durch die Gabe eines konstruktiven Feedbacks aus, wie 2008 von Wilkinson ausgeführt (Wilkinson et al. 2008).

Es gibt noch viele weitere Methoden, die Leistung von Studierenden zu erfassen und zu dokumentieren, teils auch mit recht unterschiedlichen Ansätzen und auch Zielen. Die 360-Grad-Evaluierung möchte viele Aspekte der Arbeit einer zu beurteilenden Person durch mehrere Personen aus dem Umfeld erfassen. Manche Methoden arbeiten mit Checklisten und teils sehr detailliert, andere eher grob umfassend, wie beispielsweise Global-Ratings. Nimmt man auch auf die Zeitkomponente Rücksicht, wären Log-Aufzeichnungen oder Portfolios zu erwähnen. Die verfolgten Ziele und die geprüften Inhalte bestimmen die zu wählenden Prüfungsmethoden.

2.3.4. Exkurs: Neue psychometrische Modelle

In diesem kurzen Exkurs möchte ich auf Schuwirth, seinen Kollegen Van der Vleuten und zwei seiner Arbeiten verweisen (L. Schuwirth & Van der Vleuten 2004; Schuwirth & Van der Vleuten 2006). Das besondere an diesen ist, dass sie hervorheben, dass unser aktuelles Verständnis von einer adäquaten Leistungsbeurteilung auf einer speziellen Sichtweise beruht und dass andere Sichtweisen auch denkbar und hilfreich sein können.

Die verwendeten psychometrischen Modelle haben das Denken der letzten Dekaden dominiert, Faktoren wie Reproduzierbarkeit, Validität und Effizienz waren wichtig. Aber diese Modelle kommen aus der Psychologie und aus Persönlichkeitstests, bei denen ein Item nicht zwangsweise *intrinsically meaningful* sein musste. Schuwirth beschreibt in dieser Arbeit auch die Implikationen: die Reduktion der Leistungsbeurteilung auf numerische Daten, die noch dazu den bestrafenden Aspekt über den belohnenden Aspekt heben. Itemanalysen haben oft eine so große Bedeutung, dass Items aus Prüfungsauswertungen entfernt werden, weil die Kennwerte schlecht sind, obwohl es sich nachweislich um valide Items handelt. Am Ende werden die numerischen Daten, auf die die Beurteilung reduziert wird, noch zusätzlich dichotomisiert – es läuft auf eine Entscheidung bestanden/nicht-bestanden hinaus.

Er plädiert im zweiten Teil von *Merging views on assessment* dafür, die Stärken und Schwächen der Studierenden hervorzuheben, mitzuhelfen, die individuellen Lernpfade zu optimieren und die Beurteilungsergebnisse höchst informativ zu gestalten. Es soll angestrebt werden, das ganze Bild zu sehen, wofür man eine Vielzahl an Methoden braucht. Die Ergebnisse unterschiedlicher Methoden sollen komprimiert und zusammengeführt eine professionelle Beurteilung ermöglichen – ähnlich der Zusammenführung der unterschiedlichen Informationen auf einer Fieberkurve eines Patienten.

Die Reliabilität wird dabei auf andere Weise unterstützt: Wenn man sich in der Beurteilung unsicher ist, werden weitere zusätzliche Informationen eingeholt, die das Bild vervollständigen und dann eine Entscheidung ermöglichen.

Diese andere Sichtweise, für die andere Theorien und Modelle erarbeitet werden müssen, könnte uns am Ende ermöglichen, moderne Methoden wie mini-CEX, 360-Grad-Feedback und longitudinale Formate richtig anzuwenden und zu interpretieren. Beide Autoren betonen dabei aber auch, dass beide Sichtweisen nebeneinander existieren und sich gegenseitig unterstützen können.

2.4. INHALTE – ODER WAS WIR PRÜFEN

Dieser Abschnitt soll die unterschiedlichen inhaltlichen Dimensionen beleuchten und zudem das entscheidende Gütekriterium *Validität* erklären. Eine objektive und zuverlässige Messung ist die Grundvoraussetzung, die Validität für eine glaubhafte Interpretation des Testergebnisses aber genauso unabdingbar.

2.4.1. MILLER-PYRAMIDE UND BLOOM-TAXONOMIE

Das Thema Prüfungsinhalte ist ebenso vielgestaltig wie komplex, daher möchte ich nur zwei Veranschaulichungen herausgreifen: die Miller Pyramide und die Bloom Taxonomie.

Die in Abbildung 2 dargestellte Pyramide wurde ursprünglich von Miller 1990 publiziert (Miller 1990), eine erweiterte und recht bekannte Darstellung findet sich unter anderem veröffentlicht in The Lancet (Wass et al. 2001). Sie schematisiert wichtige Facetten der klinischen Kompetenz und setzt sie zudem mit möglichen – weil passenden – Prüfungsformaten in Verbindung.

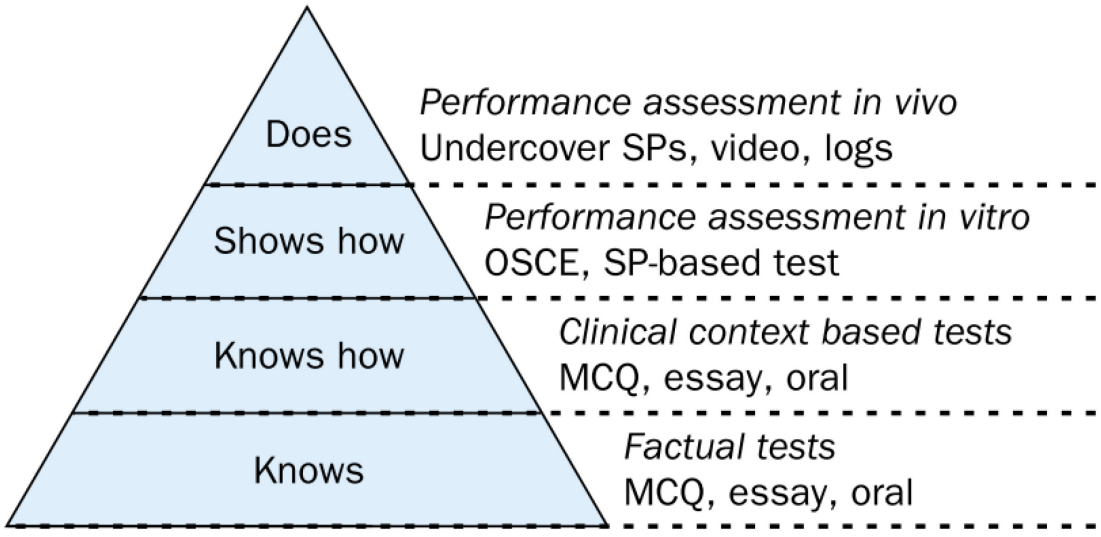

Abbildung 2: Miller Pyramide, mit den Facetten der klinischen Kompetenzen und den dazu passenden Prüfungsformaten (unverändert aus Wass et al. 2001).

Von grundlegenden Fakten, über das Anwenden von Wissen und der Fähigkeit, kontextabhängig Probleme lösen zu können, bis hin zur praktischen Betätigung und Umsetzung in der realen klinischen Welt sind sehr viele Facetten übersichtlich dargestellt. Damit wird auch recht schnell ersichtlich, dass einzeln verwendete Prüfungsformate kein umfassendes Bild der Leistung eines Studierenden liefern können (Boursicot et al. 2011).

Da wir uns in weiterer Folge mit einem schriftlichen Format beschäftigen und damit auch die kognitive Leistung dominierend sein wird, ist hier die Erwähnung von Benjamin Blooms Lernzieltaxonomie angebracht (im Original: „Taxonomy of educational objectives. Vol. 1: cognitive domain“). Allerdings verweise ich auf eine überarbeitete Form von Krathwohl (Krathwohl 2002). Die Originaltaxonomie arbeitet mit sechs Kategorien aus dem kognitiven Bereich: Wissen, Verständnis, Anwendung, sowie Analyse, Synthese und Evaluation.

Krathwohl hat dann die Originaltaxonomie um die *Wissensdimension* erweitert: Faktenwissen, konzeptuelles Wissen, prozedurales Wissen und metakognitives Wissen (wie in Abbildung 3 ersichtlich).

	The Cognitive Process Dimension					
The Knowledge Dimension	1. ***Remember***	2. ***Understand***	3. ***Apply***	4. ***Analyze***	5. ***Evaluate***	6. ***Create***
A. ***Factual Knowledge***						
B. ***Conceptual Knowledge***				X		X
C. ***Procedural Knowledge***						
D. ***Metacognitive Knowledge***						

Abbildung 3: Matrix mit den sechs Kategorien aus dem kognitiven Bereich, ergänzt um die Wissensdimensionen (unverändert aus Krathwohl 2002).

Die ersten drei Kategorien von Benjamin Bloom lauten Wissen, Verständnis und Anwendung. So würde man beispielsweise bei einer Frage, die auf reines Wiedergeben von Wissen abzielt, nach der Definition der statistischen Lagemaße *Mittelwert* und *Median* fragen. Eine Verständnisfrage würde darauf abzielen, sich eine Begründung geben zu lassen, warum in einem gegebenen Fall der Median und nicht der Mittelwert die richtige Kenngröße ist. Möchte man das Ganze als Anwendungsfrage adaptieren, könnte man die Daten (und deren Gewinnung) direkt präsentieren und abschließend nach dem sinnvollsten Lagemaß fragen.

2.4.2. VALIDITÄT

Validität, oft übersetzt mit Gültigkeit, kann als Belastbarkeit der auf Messungen beruhenden Aussagen verstanden werden (Lienert & Raatz 1998). Als Frage formuliert: Misst das Messinstrument, was es messen soll? Es gibt zahlreiche Arten von Validität, beispielsweise hat die American Psychological Association 1954 vier vorgeschlagen: Inhaltsvalidität, Konstruktvalidität und prognostische und diagnostische Kriteriumsvalidität.

Kurz umrissen, ist die Inhaltsvalidität durch Expertenurteile eher subjektiv und die Kriteriumsvalidität häufig durch das Fehlen eines geeigneten Außenkriteriums schwer zu bestimmen. Unter dem *Konstrukt* versteht man theoretische Eigenschaftsdimensionen, da man die interessierenden Eigenschaften nicht direkt erfassen und messen kann. Valide Tests lassen eine präzise und glaubhafte Interpretation der Testergebnisse zu. Beispielsweise kann aus einem Test mit 10 MC-Items über Diabetes mellitus nicht geschlossen werden, dass der Testnehmer oder die Testnehmerin ausreichende Kenntnis der Zusammenhänge in der gesamten Endokrinologie hat.

Downing hat einige Arbeiten veröffentlicht, in denen er den Validitätsbegriff noch näher ausführt und um die Begriffe Construct-Irrelevant Variance (CIV) und Construct Underrepresentation (CU) erweitert (Downing 2002; Downing & Haladyna 2004).

Bezogen auf schriftliche Tests in der Medizin wären Beispiele für CIV:

- Zu wenige Items, um das Themenfeld abdecken zu können.
- Eine verzerrte und nicht repräsentative Auswahl an Items.

Beispiele für CU wären hingegen:

- Fehlerhafte/mangelhafte Items.
- Unverständliche Items und mehrdeutige Formulierungen.
- Zu schwere, zu leichte oder nicht trennscharfe Items.
- *Schwindeln* oder Bekanntheit der Items.
- *Auf die Prüfung ausgerichtete Lehre (Teaching-to-the-test),* aber auch
- *Testfertigkeit (testwiseness).*

Abschließend muss noch die Augenscheinvalidität (face-validity) erwähnt werden, da sie aus Studierendensicht eine wichtige Form darstellt. Wenn diese Form der Schein-Validität vorliegt, sehen die Studierenden die Frage als sinnvoll und vor allem wichtig für die spätere Zukunft als Arzt oder Ärztin.

Diese Form kann die Akzeptanz einer Frage oder einer Prüfung positiv beeinflussen (Krebs 2004). Allerdings betont Downing, dass es sich hierbei um keine legitime Form von Validität handelt, obwohl er diese Eigenschaft als wichtig ansieht (Downing & Haladyna 2004). Es ist also keine anerkannte Form der Validität, sondern eher als sinnvolle Eigenschaft im Sinne der Akzeptanz einer Prüfung zu sehen.

2.5. COMPUTERBASIERTES PRÜFEN

Die Art und Weise, wie Computer beim Prüfen unterstützen können, ist vielfältig. In allen Phasen kann der Computer unterstützen: In der Vorbereitungs-, Durchführungs- und Nachbearbeitungsphase. Im folgenden Abschnitt werden zwei Arbeiten zusammenfassend wiedergegeben, die einen guten Überblick zu diesem Thema geben (Cantillon et al. 2004; Frey 2006).

Die Unterstützung durch Computer, oder allgemeiner formuliert die elektronische Datenverarbeitung, hat beim Einlesen und Auswerten von Antwortbögen Papier-basierter Prüfungen begonnen. Mittlerweile ist aber die Durchführung der Prüfung, also die Dateneingabe über eine entsprechende Benutzeroberfläche möglich und wird häufig genutzt. Egal ob die Prüfungen formativen oder summativen Charakter haben, ob sie als Webanwendung funktionieren oder aus Sicherheitsgründen isoliert genutzt werden – die Vorteile einer elektronischen Darbietung sind vielfältig. Multimediale Inhalte, ob bewegte Bilder oder zusätzliche Vertonung, können leichter und hochwertiger genutzt werden, unmittelbares Feedback für den Lernenden (im Rahmen eines formativen Szenarios) ist möglich und rasche Auswertungen ermöglichen auch adaptive Prüfungsverfahren.

Auf die Vorbereitungsphase hinweisend, kann auch die Fragen-, Aufgaben- oder Item-Verwaltung elektronisch und vernetzt erfolgen, was Reviewprozesse und automatisiertes Zusammenstellen einer Prüfung erleichtert. In der Nachbearbeitungsphase sind automatisierte Berechnungen und damit verknüpfte Prozesse möglich.

Die einzigen deutlicheren Schwierigkeiten könnten technischer Natur sein, da für Prüfungen mit großer Teilnehmerzahl auch entsprechend viele Computer oder Eingabe-Schnittstellen zur Verfügung stehen müssen. Auch wenn mehrere Prüfungsdurchgänge hintereinander möglich sind, kann es sicherheitstechnische Schwierigkeiten geben: die Verwendung anderer Prüfungsitems bei nachfolgenden Durchgängen wäre ratsam. Die Verwendung eigener Computer/Laptops durch Studierende oder sogar eine dezentrale Abwicklung ist dabei nur für formative Szenarien denkbar.

3. Prüfungsformat Multiple-Choice

3.1. Definition und Aufbau

MC-Fragen oder MC-Items werden auch als Mehrfachauswahl-Fragen bezeichnet und die Prüfungsdurchführung mit solchen als Antwort-Wahl-Verfahren. Ihr Hauptmerkmal ist das Vorhandensein mehrerer vorformulierter Antwortalternativen. Damit handelt es sich – wie bereits erwähnt – um einen geschlossenen Fragentyp, im Gegensatz zum offenen Fragentyp, bei dem Studierende eigene Antworten formulieren müssen.

Der Terminus MC-Frage umfasst jedoch viele unterschiedliche Formate, die eben nur die Tatsache gemein haben, dass vorgefertigte Antwortalternativen zur Verfügung stehen. Angenommen, man bietet mehrere Antwortalternativen an, so können entweder eine oder mehrere Alternativen richtig sein. Zudem kann die Anzahl der richtigen Antwortalternativen entweder bekannt, oder nicht bekannt sein.

Der Aufbau einer MC-Frage ist prinzipiell immer derselbe: eine Problemstellung (oft auch als Vignette bezeichnet) bildet zusammen mit der Fragestellung den Fragenstamm, darauf folgen die Antwortoptionen (kurz AWO), wobei die falschen Antwortoptionen als Distraktoren bezeichnet werden. Warum die fachlich inkorrekten Antwortoptionen Distraktoren heißen, wird klar, wenn man sich die Bedeutung des englischen Worts *distract* ansieht: diese Optionen sollen von der korrekten Antwortoption ablenken. In der englischsprachigen Fachliteratur wird der Stamm als *stem* und die Fragestellung als *lead-in* bezeichnet, die Antwortalternativen als *options*, wobei die korrekte als *key* und die inkorrekten – wie bereits erklärt – als *distractors* bezeichnet werden.

MC-Fragen bieten den Vorteil, dass sie objektiver sind und ein effizientes Prüfungsinstrument darstellen, was vor allem bei einer hohen Zahl an Teilnehmenden interessant ist. In gleicher Weise ist auch die Auswertung ressourcenschonend, speziell in der heutigen Zeit, in der mit Hilfe des Computers ausgewertet wird. Vorausgesetzt, es werden genügend Fragenitems verwendet und diese sind sauber, eindeutig und ohne Lösungshinweise formuliert, hat das MC-Format auch eine hohe Reliabilität im Vergleich zu anderen Formaten vorzuweisen (wie in Abbildung 4 ersichtlich).

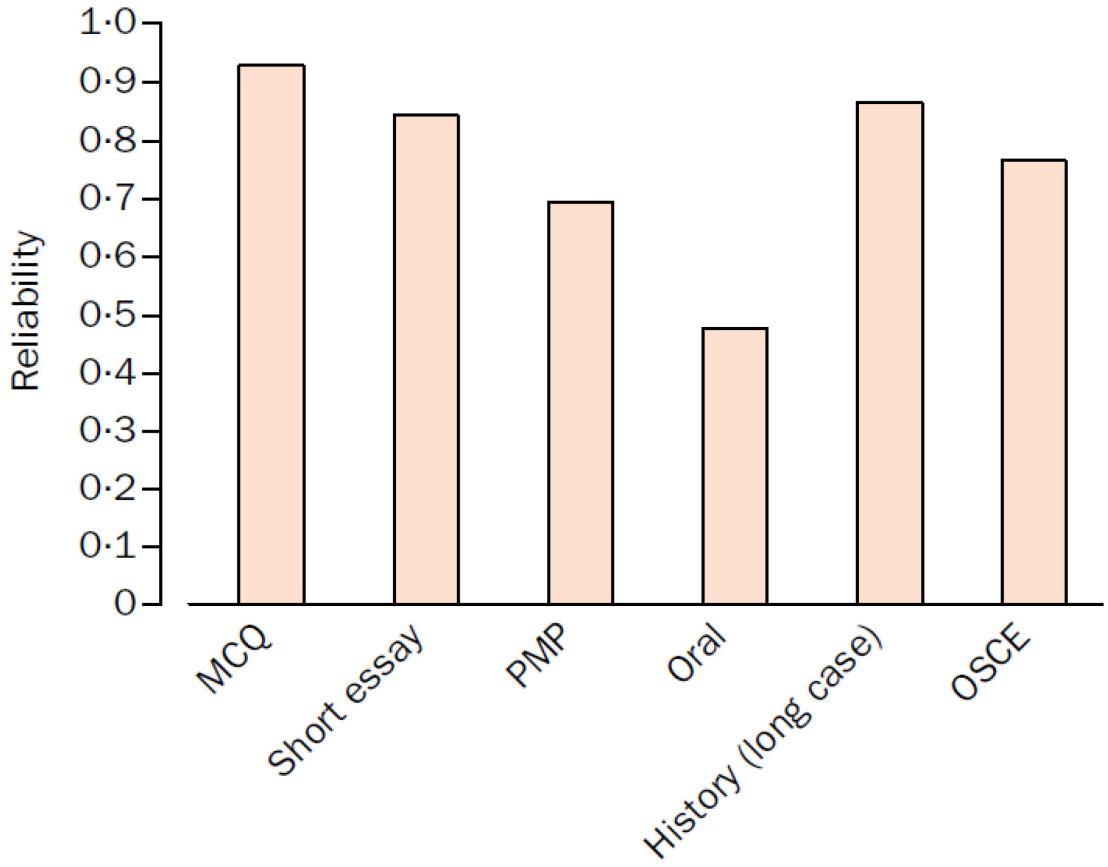

Abbildung 4: Grober Reliabilitätsvergleich unterschiedlicher Prüfungsformate, angepasst an eine Testdauer von vier Stunden (unverändert aus Wass et al. 2001)

Das MC-Format hat aber auch einige Nachteile, viele davon durch das Vorhandensein von *vorformulierten* Antwortoptionen bedingt. Naturgemäß hat dieses Format bereits eine Ratewahrscheinlichkeit über null, beim klassischen MC-Typ mit fünf Optionen liegt diese bereits bei 20%, da aber nicht alle Distraktoren gleich plausibel klingen, steigt diese schnell weiter an (Kubinger 2005). Daneben gibt es noch zahlreiche andere Aspekte wie beispielsweise Teilwissen, fehlende Eindeutigkeit bei der Formulierung oder die immer mitgetesteten linguistischen Fähigkeiten. Der wichtigste Punkt, der abschließend erläutert werden soll, betrifft die Testfähigkeit oder im Original *testwiseness*.

Studierende, die mit schriftlichen MC-Prüfungen bestens vertraut sind, nehmen beispielsweise ungewollte Lösungshinweise instinktiv auf und verbessern damit ihre Ratewahrscheinlichkeit deutlich. Dies trifft vor allem auf die Studierenden zu, die in ihrer schulischen und studentischen Laufbahn bereits sehr viel mit MC-Prüfungen zu tun hatten. Millman, Bishop und Ebel haben diese 1965 folgendermaßen beschrieben: „*Subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score. Test wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures*“ (Millman et al. 1965). Wie geschickt manche Studierenden mit MC-Fragen umgehen, zeigt beispielsweise auch eine deutsche Arbeit aus dem Jahr 2005 (Schulze et al. 2005).

Daneben gibt es noch Arbeiten, die zeigen, dass sich das nachträgliche Ändern der ausgewählten Antwortoption stets positiv, also im Sinne einer Verbesserung, auswirkt (Fischer, Herrmann, et al. 2005; Bauer et al. 2007; Fabrey & Case 1985).

Die Entwicklung verschiedener Formate und Methoden zur objektiven schriftlichen Prüfung hat bereits vor dem 21. Jahrhundert begonnen und sich Schritt für Schritt verändert. Das *National Board of Medical Examiners* hat schließlich die vorgeschlagenen Fragetypen mit Buchstaben gekennzeichnet - diese sind heute noch großteils gültig und vor allem gebräuchlich. Es lassen sich grundsätzlich zwei große Gruppen von Frage-Typen unterscheiden: Beste-Antwort-Typen und Richtig-Falsch-Typen.

Inhaltlich werden MC-Fragen oft in zwei Gruppen unterteilt, in das *Beste-Antwort-Format* und das *Richtig-Falsch-Format*. Ersteres wird bei anwendungsorientierten und praxisnahen Fragestellungen eingesetzt, bei der mit Wahrscheinlichkeiten gearbeitet wird und die einer Graustufenabwägung entsprechen. Zweiteres wird eingesetzt, wenn es um eine eindeutig zu beantwortende Frage geht, bei der eine Trennung zwischen korrekt und inkorrekt eindeutig im Sinne einer Schwarz-Weiß-Entscheidung möglich ist. Dies trifft auf viele Faktenfragen in Grundlagenbereichen zu.

3.2. Gruppe der Beste-Antwort-Fragen

Der *klassische* und eindeutig zu bevorzugende Fragentyp heißt A-positiv – auch *Best-of-five* genannt – und besteht aus einem Fragenstamm und fünf Antwortoptionen, von denen eine korrekt ist.

Die Fragestellung soll positiv, aber auch kurz, prägnant und eindeutig formuliert sein. Die Antwortoptionen sollen nach Möglichkeit homogen und ebenfalls prägnant formuliert sein. Diese drei Sätze stellen jedoch nur die minimale Variante einer Empfehlung für das Schreiben hochwertiger Fragenitems dar, diesem Thema ist ein eigenes Kapitel gewidmet (4.2). Bei Vorkommen einer Problemstellung soll diese realitätsnahe und positiv formuliert sein und alle für die Lösung notwendigen Informationen enthalten (Schuwirth 1999). Aufgrund solider Evidenz bezüglich seiner Qualität sollte dieser Typ bei jeder Prüfung klar dominieren (Case & Swanson 1998).

Für das Kennen wichtiger Ausnahmen eignet sich der Typ-A-negativ (Krebs 2004). Dieser Fragentyp gleicht in allen Punkten der Typ-A-positiv-Frage, die Frage ist lediglich umgekehrt formuliert und fragt daher nach der *Ausnahme*. Rene Krebs schreibt diesem Typ vergleichbare Item-Schwierigkeit und Item-Trennschärfe zu, sieht allerdings eine reduzierte Validität. Die Antworten sind dennoch positiv formuliert, um Doppelnegationen zu vermeiden.

Die Negation muss durch Fettdruck, Unterstreichung oder Großbuchstaben – auf alle Fälle einheitlich – hervorgehoben werden!

Daneben sind noch weitere MC-Typen erwähnt, die allerdings nur Abwandlungen und Varianten der oben genannten darstellen. So ist beispielsweise der Typ-B nur eine Kombination aus mehreren Typ-A-Fragen mit denselben fünf Antwortoptionen. Der Typ-R gleicht grundsätzlich ebenfalls dem Typ-B, die Anzahl der Antwortoptionen kann hier jedoch bis zu 26 betragen (Auflistung mit den Buchstaben A-Z). Der Vorteil dieses Typs nimmt auf die Art und Weise der Beantwortung Bezug, das Suchen der potentiell richtigen Lösung wird bei bis zu 26 Optionen unökonomisch, daher erhofft man sich von den Teilnehmenden ein selbstständiges Erarbeiten der Lösung, welche anschließend in der Liste von direkt aufgesucht wird.

Der Fragentyp PickN – welcher 1996 von Ripkey, Case und Swanson (Ripkey et al. 1996) vorgestellt wurde – gleicht ebenfalls der Typ-A-Frage, allerdings mit dem Unterschied, dass mehr als eine Antwortoption richtig sein kann. Die Anzahl der korrekten und daher zu wählenden Antwortoptionen sollte dabei angegeben werden. Dieser Typ ist dann zu bevorzugen, wenn es um Graustufenabwägungen geht, für einfachere richtig-falsch-Entscheidungen wäre der nachfolgend erklärte Typ-Kprim zu bevorzugen. Bedenken sollte man jedoch die Sinnhaftigkeit von Teilpunkten, vor allem wenn es mehr als zwei korrekte Optionen gibt, was auch Bauer 2011 (Bauer et al. 2011) bestätigt hat.

3.3. GRUPPE DER RICHTIG-FALSCH-FRAGEN

Zur zweiten Gruppe zählt der nicht empfohlene Typ-K, auch als *Kombinations-frage* bekannt. Dieser beginnt mit einer Frage mit vier Aussagen zum Fragenstamm, die Antwortoptionen sehen dann typischerweise wie folgt aus: „1 und 3 sind richtig“, „2 und 4 sind richtig“, „1,2 und 3 sind richtig“ oder „nur 4 ist richtig“ (wie in Abbildung 5 ersichtlich). Die Problematik bei diesem Typ stellen die Kombinationsmöglichkeiten dar, mit Teilwissen können Teilnehmende schnell zu einer *50-50* Ratewahrscheinlichkeit kommen (Krebs 2004).

Albanese fasst es in seiner Arbeit folgendermaßen zusammen: der Fragentyp-K ist einfacher, die Reliabilität sinkt und Typ-K ist anfällig für Mängel (Albanese 1993).

Übergeordneter Fragentext…

Aussage 1
Aussage 2
Aussage 3
Aussage 4

(A) Aussage 1, 2 und 3 sind richtig.
(B) Aussage 1 und 3 sind richtig.
(C) Aussage 2 und 4 sind richtig.
(D) Nur Aussage 4 ist richtig.
(E) Alle Aussagen sind richtig.

Abbildung 5: Grundlegender Aufbau einer Frage des Typs-K. Die Kombinationsmöglichkeiten einzelner Aussagen repräsentieren die Antwortoptionen (A) bis (E).

Ähnlich problematisch ist der Typ-RF (Richtig/Falsch) da hier ebenfalls eine 50-50 Ratewahrscheinlichkeit vorliegt. Man findet gelegentlich auch den „*Typ-RF?*“ (sic!) (Richtig/Falsch/Weiß nicht), der dem Typ-RF bis auf die Weiß-nicht-Option gleicht. Eine Überlegung geht dahingehend, bei Richtig und Falsch einen Punkteabzug bei falscher Beantwortung durchzuführen, jedoch nicht bei „Weiß nicht“.

Die mehrfache Entscheidung richtig-falsch nennt sich Typ-Kprim oder englisch-sprachig MTF (multiple true-false), dabei müssen vier Aussagen – welche auf eine Frage folgen – getrennt voneinander auf ihre Richtigkeit überprüft werden. Der Fragentyp ist in der folgenden Abbildung 6 grob skizziert.

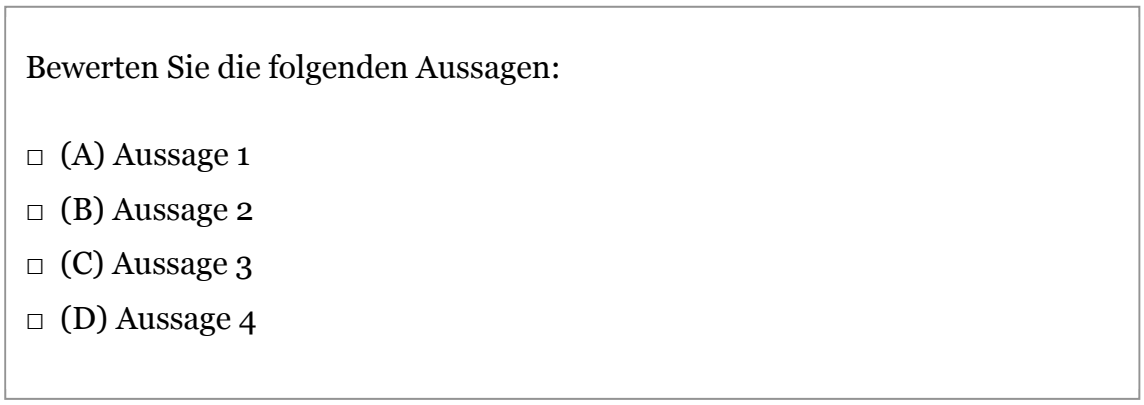

Abbildung 6: Grundlegender Aufbau einer Frage des Typs-Kprim bzw. MTF (Multiple True False). Alle vier Aussagen sind getrennt voneinander auf ihre Korrektheit zu überprüfen. Jede Kombination von *alle richtig* bis *alle falsch* ist möglich.

Eindeutig richtige oder eindeutig falsche Aussagen sind Voraussetzung, dieser Typ sollte nicht missbraucht werden, um heterogene Aussagen zusammen zu ziehen. Auch hier sind Teilpunkte – beispielsweise ein halber Punkt bei drei von vier richtigen Antworten – sinnvoll.

In Summe gibt es zahlreiche Fragentypen, die sich teilweise auch nur wenig unterscheiden. Deshalb ist es wichtig, sich mit wenigen Typen intensiv zu beschäftigen und das Fragenschreiben mit diesen auch eingehend zu üben. Krebs beschreibt diesen Gedanken folgendermaßen (Krebs 2004):

> *„Zunehmend wurde aber erkannt, dass es unter den Aspekten der Herstellung, der Beantwortung und der Messqualität besser ist, sich auf wenige Typen zu beschränken.“*

3.4. ANWENDUNGSORIENTIERTE FRAGEN

Schuwirth hat 2004 mit Hilfe der Eigenschaften Stimulus und Response eine hilfreiche Einteilung geschaffen, um schriftliche Formate besser überblicken zu können (L. W. T. Schuwirth & Van der Vleuten 2004). Früher hat man von *Case-based* oder *Short-Cases* gesprochen, Schuwirth zählt aktuell nur Extended-Matching Fragen und Key-Feature Fragen auf. Unabhängig von der Tatsache, wie solche Fall-orientierten Fragen betitelt werden, muss wiederholt werden, dass die Problemlöse-Fähigkeit nicht generisch und einheitlich ist, sondern im höchsten Grad fach- und themenspezifisch ist. Deshalb sollten viele, aber kurze Fallpräsentationen verwendet werden (Schuwirth 1999; Schuwirth & Verheggen 2001).

Bei *Extended-Matching-Fragen* (kurz EMQ) gibt es eine Liste von 10-20 Antwortoptionen. Danach werden mehrere kurze Fallvignetten präsentiert und für alle gelten dieselben Antwortoptionen. Dabei gibt es Optionen, die für eine Vignette passend sind, vielleicht welche, die für mehrere Vignetten passen, andere Optionen passen bei keiner Vignette. Durch die Anzahl an Optionen und die Verwendung bei mehr als einer Vignette soll das Finden von ungewollten Lösungshinweisen unterbunden werden (Case & Swanson 1993). Auch Beullens bestätigt, dass man mit Hilfe von Extended-Matching Fragen klinisches Wissen und klinisches Denken beurteilen kann (Beullens et al. 2002; Beullens et al. 2005).

Bei *Key-Feature-Fragen* (kurz KF) werden kurze Fälle mit klinischen Zeichen, Symptomen und Untersuchungsergebnissen beschrieben, gefolgt von wenigen Fragen zu diesem Fall. Page und Bordage, stellvertretend für das Medical Council of Canada, haben es 1995 in ihrer Arbeit beschrieben, da unter anderem die Evidenz für die zuvor verwendeten PMPs (Patient-Management-Problems) nicht vorhanden war (Page & Bordage 1995). Alle Fragen zielen dabei auf wichtige *Schlüsselentscheidungen* ab.

Damit kann die Prüfungszeit pro Frage kurz gehalten werden, wodurch viele (bis zu 30) Fälle pro Stunde Prüfungszeit abgefragt werden können. Es hat sich herausgestellt, dass dieses Format valide und reliabel Problemlösefähigkeit abfragen kann (L. Schuwirth & Van der Vleuten 2004). Auch im deutschen Sprachraum sind einige unterstützende Studien dazu veröffentlicht worden (Fischer, Kopp, et al. 2005; Kopp & Möltner 2006).

Ergänzend seien *Modified-Essay Fragen* (kurz MEQ) erwähnt, welche von Palmer und Devitt beschrieben (Palmer & Devitt 2007) wurden. Letzten Endes sind sich Modified-Essay-Fragen und MC-Fragen zu ähnlich, wie die beiden Autoren festhalten. Zudem ist auch die Konstruktion dieses Fragentyps nicht unproblematisch.

Abschließend ein Format, welches Schuwirth als hybrides Format beschreibt: der *Script-Konkordanz-Test*. Dieser fragt nach einer Wahrscheinlichkeit einer Diagnose zu einer Fallbeschreibung, bei der ein zusätzliches klinisches Zeichen oder klinisches Symptom angeführt ist. Dieses Format wurde von B. Charlin im Rahmen einer Dissertation an der Universität Maastricht 2002 beschrieben und basiert auf Überlegungen zu Problemlösefähigkeit und Expertise. Ziel ist es, die Stärke von semantischen Netzwerken und das Vorhandensein von *illness-scripts* zu messen – diese als Vorannahme vorausgesetzt. Die Idee wurde ebenso im Jahr 2000 (Charlin et al. 2000) und 2005 (Sibert et al. 2005) näher beschrieben.

4. Erstellung und Verwendung von Multiple-Choice-Fragen

4.1. Prozess der Fragen-Erstellung

"*Effective item writing is both art and science*", wie es Downing treffend beschreibt (Downing 2006). Dieser Abschnitt konzentriert sich auf die Erstellung oder die Konstruktion von MC-Fragen. Vor dem eigentlichen Erstellen neuer MC-Fragenitems sollte man sich Zeit für einige Vorbereitungen nehmen. Das bedeutet, alle fachspezifischen Unterlagen bereitzulegen, die man während der Erstellung benötigen könnte: Vorlesungsfolien, Lehrbücher, Lernziel- und Schlagwortkataloge, sowie *Blueprints* (gewichtete Rasterung der Prüfungsinhalte).

Daneben benötigt man eventuell Leitfäden und Ratgeber zur MC-Fragenerstellung, wie beispielsweise der Leitfaden von Smolle (Smolle 2008), sowie Checklisten, anhand derer man die erstellten Fragen überprüfen kann. Zusätzlich können auch Formulare und Vorlagen hilfreich sein, in die man die entsprechenden Fragen-Inhalte einträgt. Zudem sollte man sich als Autorin oder Autor mit einer bestehenden Fragen-Verwaltungssoftware oder bestehenden Peer-Review-Systemen vertraut machen. Gleiches gilt für den Fall, dass mit Lernplattformen, wie beispielsweise Moodle[10], ILIAS[11] oder OLAT[12] gearbeitet wird.

[10] http://moodle.org (abgerufen am 12. August 2015).
[11] http://www.ilias.de (abgerufen am 12. August 2015).
[12] http://www.olat.org (abgerufen am 12. August 2015).

Es gibt zahlreiche Möglichkeiten, an die MC-Fragenerstellung heranzugehen – zahlreiche Aspekte, die variiert werden können: Wer arbeitet an der Fragenerstellung und wer arbeitet mit? Wie arbeitet man, mit Papier und Stift, am Computer oder am Tablet-PC? Bringt das Erstellen neuer Fragen Vor- oder Nachteile im Vergleich zum Überarbeiten von Altfragen?

4.2. ANFORDERUNGEN AN MC-FRAGEN

Von den folgenden Richtlinien und Empfehlungen stützen sich viele auf wissenschaftliche Studienergebnisse, andere wiederum bilden eine Art Expertenkonsens ab. Es gibt unzählige Quellen, wenn es um Empfehlungen und Richtlinien geht, die die Fragenitem-Erstellung unterstützen und leiten sollen.

Ich beginne mit dem Standardwerk des *National Board of Medical Examiners*, geschrieben von Case und Swanson (Case & Swanson 1998). Ich möchte dabei auf den zweiten Teil eingehen und die wichtigsten Punkte für das Schreiben von Fragen zusammenfassen:

MC-Fragen sollen fokussiert sein, am besten die Anwendung von Wissen abfragen und über eine klar formulierte Frage und homogene Antwortoptionen verfügen. Unfokussierte Fragen wie „*…welche der folgenden Aussagen…*“ sind zu vermeiden.

Im ersten Teil beschreiben Case und Swanson Aspekte, die zu einer unnötigen Steigerung des Schwierigkeitsgrades oder auch zur Verwirrung der Teilnehmenden führen können: lange und komplexe Fragestellungen, uneinheitlich präsentierte numerische oder sprachliche Ausdrücke, vage und unsichere Angaben, speziell Häufigkeitsadverbien, Antwortoptionen wie „keine der oben genannten“ oder „alle der oben genannten“ oder einfach *tricky items*.

Ein spezielles Kapitel sind *ungewollte Lösungshinweise*, in der englischsprachigen Literatur oft als *cues* bezeichnet, etwas allgemeiner formuliert auch als *flaws*. Diese sind eng mit der Testfähigkeit (*testwiseness)* verknüpft, da Teilnehmende, die diese Eigenschaft mitbringen, in der Lage sind, eben solche Hinweise schnell und gut deuten zu können. Folgende Lösungshinweise werden durch Case und Swanson beschrieben: grammatikalische und logische Hinweise, absolute Terme oder Aussagen, auffällig lange richtige Antwortoptionen, Wortwiederholungen, sowie der Konvergenz-Cue (dieser ist in Abbildung 7 anhand eines Beispiels erklärt).

Die Abkürzung USL heißt ausgeschrieben…

(A) United States Laboratories
(B) Uniform Source Language
(C) Uniform Source Locator
(D) Uniform Starting Label
(E) Unique Spaceship Locator

Abbildung 7: Der Konvergenz-Cue anhand eines ausgedachten Beispiels, welches von Krebs übernommen und leicht adaptiert wurde (Krebs 2008).

Im zuvor erwähnten zweiten Teil werden zahlreiche praktische Fragenbeispiele gebracht, die alle Mängel und ihre Korrekturen erläutern sollen. Erwähnen möchte ich abschließend die empfohlene Textverteilung bei MC-Fragenitems. Wie in der nachfolgenden Abbildung 8 ersichtlich, kann der Fragenstamm, also die Problem- und Fragestellung länger formuliert sein, solange die Antwortoptionen prägnant formuliert sind.

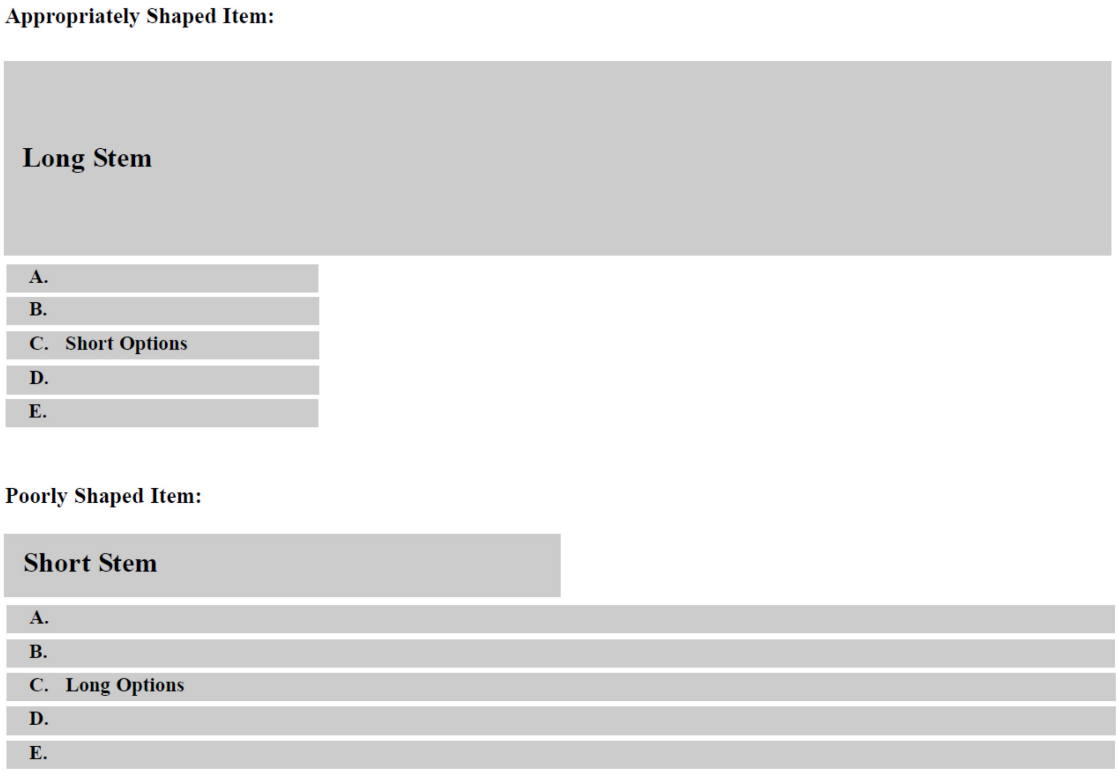

Abbildung 8: Übersicht über die Verteilung der *Textlast* bei der Erstellung von MC-Fragenitems (unverändert aus Case & Swanson 1998).

Parallel dazu und bestätigend möchte ich die Schulungsmaterialien von Krebs von der Universität Bern umreißen (Krebs 2004). Bevor er in seiner Arbeit auf die gängigsten Fragetypen eingeht, listet er Aspekte auf, die entweder die Gültigkeit oder die Zuverlässigkeit einer Frage und letzten Endes der Prüfung verbessern. Für eine entsprechende Gültigkeit fordert er relevante Frageninhalte, vor allem in Hinblick auf zukünftige Anforderungen, ein passendes Anspruchsniveau, auch bezüglich der Kategorien Wissen, Verstehen und Anwenden, sowie fokussierte Fragen mit einer eindeutig besten Lösung. Vom Schwierigkeitsgrad angemessene Fragen, die sprachlich einfach und klar formuliert sind und keine Lösungshinweise enthalten, unterstützen die Reliabilität.

Abschließend soll wiederholt auf die Relevanz einer Frage eingegangen werden, da Rene Krebs diese extra erläutert. Relevante Items ergeben sich vor allem dann, wenn es um Themen geht, die…

- …häufig in der alltäglichen Praxis vorkommen,
- …gravierende Folgen nach sich ziehen, sollte ein Fehler passieren,
- …mit verbreiteten Fehlmeinungen verknüpft sind, oder
- …für die späteren Lehrinhalte und deren Verstehen entscheidend sind.

4.2.1. Richtlinien und deren Validität

In diesem Abschnitt werden zwei umfassende und grundlegende Arbeiten vorstellen. Die erste Arbeit – eine Meta-Analyse – hat Rodriguez von der Michigan State University 1997 publiziert (Rodriguez 1997):

Am Beginn dieser Arbeit stellt er sieben intensiv bearbeitete und ausgewählte Regeln vor, die bei der Fragenerstellung zu beachten sind. Dabei zitiert er Haladyna (Thomas M Haladyna & Downing 1989):

1. Vermeiden des komplexen MC-Formats Typ-K.
2. Der Fragenstamm soll als Frage formuliert sein.
3. Dieser Stamm soll positiv formuliert sein.
4. Es sollen so viele Distraktoren wie möglich beigefügt werden.
5. Vermeiden (oder seltenes Verwenden) der Optionen „alle der oben genannten“ und „keine der oben genannten“.
6. Die Länge der Antwortoptionen sollte (ungefähr) gleich sein.

Er verweist ein weiteres Mal auf dieses Autorenduo (Thomas M. Haladyna & Downing 1989), um den Grad des Konsenses der Experten zu beschreiben. Die Vermeidung des komplexen Fragetyps-K wird nicht von allen Expertinnen und Experten gleich gesehen (in etwa gleich viele Befürworter wie Gegner), die Vermeidung einer negativen Formulierung in der Frage allerdings schon, hier spricht sich die Mehrheit für diese Regel aus.

Michael Rodriguez geht im Anschluss alle ausgewählten Regeln Schritt für Schritt durch und fasst seine Rechercheergebnisse zusammen. Die Vermeidung des Fragentyps-K wird nicht sehr oft erwähnt, auch er sieht den fehlenden Konsens, er beschreibt aber, dass sechs von sieben Arbeiten von einer erhöhten Schwierigkeit dieser Fragen ausgehen. Er verweist aber auf eine Arbeit, die beschreibt, dass Lösungshinweise bei diesem Fragentyp den Schwierigkeitsgrad deutlich reduzieren (Albanese 1982; Albanese 1993).

Bei der Regel um die Vermeidung von negativen Formulierungen sieht er eine breite Zustimmung. Andererseits beschreiben nur zwei von sechs Arbeiten einen Anstieg im Schwierigkeitsgrad (Tamir 1993; Cassels & Johnstone 1984).

Die zweite erwähnte Arbeit ist von Haladyna. Er hat mehrere Arbeiten zum Thema Richtlinien für das Schreiben von MC-Fragen veröffentlicht, unter anderem auch eine Taxonomie (Thomas M Haladyna & Downing 1989), eine Validierung derselben (Thomas M. Haladyna & Downing 1989) und einen Review im Jahre 2002 (Haladyna & Downing 2002) – auf letzteren möchte ich hier verweisen.

Er publiziert in dieser Arbeit eine Auflistung von 31 Regeln, mit Regel Nr. 9 („avoid type K“) und Regel Nr. 17 („avoid negatives“) auch zwei sehr interessante Empfehlungen. Bei Regel Nr. 17 beschreibt er einen fehlenden Konsens, der komplexe Fragentyp-K hat drei zitierte *Gegner*. Eine zitierte Arbeit spricht sich für ein Ersetzen mit dem Typ MTF (multiple true-false) aus (Frisbie 1992), Albanese spricht von einem *poor type*(Albanese 1993), die dritte Arbeit sieht eine erhöhte Schwierigkeit bei diesem Typ (Nnodim 1992).

4.2.2. Anzahl der Distraktoren

Eine wichtige und immer wieder gestellte Frage zielt auf die sinnvolle Anzahl an Distraktoren ab. Es gibt einige unterschiedliche spontan argumentierbare Antworten: Beim Fragentyp-A positiv, mit dem Beinamen *best-of-five*, sollen es die geforderten vier sein. Durch die sinkende Ratewahrscheinlichkeit könnte man auch mehr Distraktoren plausibel argumentieren. Auf der anderen Seite kann man diese nicht erzwingen, es ist daher besser, man hat nur drei gute Distraktoren als eine vierte wenig sinnvolle Variante.

Ich möchte hier Arbeiten anführen, die einheitlich festhalten, dass die optimale Anzahl an Antwortoptionen drei (sic!) ist (Rodriguez 2005; Vyas & Supe 2008). Laut Rodriguez können bei einer Prüfung besser mehr Items mit drei Optionen als Items mit mehr Optionen gestellt werden, was den Inhalt besser abdeckt, ohne die psychometrische Qualität des Ergebnisses zu beeinträchtigen. Auch Vyas und Supe sehen die Aussagekraft nicht beeinträchtigt, halten darüber hinaus fest, dass diese kürzeren Varianten schneller erstellt werden können. Eine jüngere Studie mit derselben Kernaussage ist auch die Arbeit von Schneid (Schneid et al. 2014).

4.2.3. Art und Qualität der Distraktoren

In diesem Abschnitt soll auf zwei Besonderheiten eingegangen werden: Einerseits geht es um das praktische Erstellen von Distraktoren, vor allem auch bei negativen Fragenitems. Andererseits geht es um die Heterogenität und das Vorhandensein mehrerer Dimensionen in einem Fünfer-Satz an Antwortoptionen.

Das Finden von guten Distraktoren ist eine schwierige Angelegenheit, da man inhaltlich eindeutig falsche Optionen benötigt, die aber plausibel klingen und von den Studierenden ernsthaft überlegt werden. Wenn man Fragen vom Typ-A positiv schreibt, hat man diese Schwierigkeit – oder *Gradwanderung* – viermal, wenn man hingegen Typ-A negative Fragen schreibt, nur einmal. Dieser Vergleich ist in Abbildung 9 anschaulich dargestellt.

Frage Typ-A positiv	Frage Typ-A negativ
Korrekte Aussage (zu wählen)	Inkorrekte Antwort (zu wählen)
Inkorrekte Aussage (Distraktor)	Korrekte Antwort (Distraktor)
Inkorrekte Aussage (Distraktor)	Korrekte Antwort (Distraktor)
Inkorrekte Aussage (Distraktor)	Korrekte Antwort (Distraktor)
Inkorrekte Aussage (Distraktor)	Korrekte Antwort (Distraktor)

Abbildung 9: Übersicht über den Aufbau positiv und negativ formulierter Fragen aus den einzelnen Antwortoptionselementen.

Ein wichtiges Kriterium ist die Homogenität der Antwortoptionen, Optionen aus unterschiedlichen Dimensionen eines Themengebietes sind zu vermeiden. Wenn man beispielsweise nach einer Diagnose fragt, soll man fünf Diagnosen auflisten. Bei Medikamenten, Krankheitserregern, Symptomen oder Antworten auf Patientenfragen ist es dasselbe – immer fünf homogene zusammengehörige Optionen.

Wie rasch man diese grundlegende Idee missachten kann, zeigen *Aussagen-basierte Fragen*. Diese sind nach dem Schema „welche Aussage ist richtig“ oder „welche Aussage ist falsch“ aufgebaut und verleiten extrem zu heterogenen Antwortoptionen. Die Frage ist breit, der Kontext ebenfalls weit gefasst und der Fokus ebenfalls unscharf. Da man beim Zusammenstellen der einzelnen Aussagen nicht so wählerisch und präzise sein muss, findet man leichter Optionen. Diese führen dann allerdings zu schwereren Fragenitems, da unterschiedliche Dimensionen miteinander verglichen werden müssen.

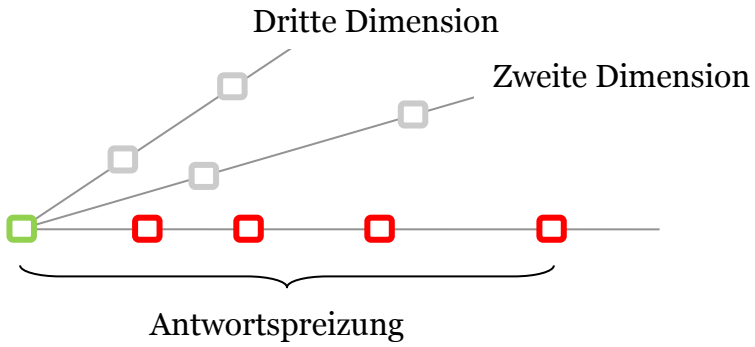

Abbildung 10: Schematische Darstellung des Verhältnisses der einzelnen Antwortoptionen, der unterschiedlichen Dimensionen und ergänzend der Antwortspreizung.

In der Abbildung 10 wurde noch ein weiteres Detail eingefügt, welches beim Erstellen von falschen Antwortoptionen entscheidend ist, die Antwortspreizung. Dieser Term zielt darauf ab, wie weit die Optionen auseinanderliegen und beeinflusst den Schwierigkeitsgrad einer Frage enorm. Je weiter die Optionen auseinanderliegen, desto leichter ist das Item, je enger beisammen, desto schwerer ist es.

4.2.4. Anwendungsorientierte MC-Fragen

Schuwirth hat 2004 in seiner Arbeit die Einteilung aller MC-Fragetypen anhand von Stimulus und Response propagiert (L. W. T. Schuwirth & Van der Vleuten 2004). Demnach unterscheidet er beim Stimulus Fragen mit und ohne Kontext, beim Antwortverhalten offene Fragen und MC-Fragen. Kontextfreie Fragen prüfen vorwiegend Faktenwissen, welches für das Problemlösen wichtig sind, allerdings nur einen Aspekt darstellen (Boshuizen & Schmidt 1992; Hager & Gonczi 1996).

Kontext-Fragen initiieren einen ganz anderen Denkprozess, bei dem das Abwiegen und Vergleichen von Informationselementen dominiert. Moderne Fragetypen, bei denen solche Fälle oder Vignetten präsentiert werden, sind entweder Extended-Matching-Fragen (Case & Swanson 1993) oder Key-Feature-Fragen (Page & Bordage 1995).

Grundsätzlich gelten für anwendungsorientierte MC-Fragen dieselben Richtlinien, allerdings kommt hier auch das Schreiben und Formulieren der Fälle und Vignetten hinzu. Wie es in den Empfehlungen des *National Board of Medical Examiners* steht (Case & Swanson 1998), können solche Patienten-Vignetten aus einer Vielzahl von Elementen bestehen:

- Alter und Geschlecht
- Ort einer fiktiven Kontaktaufnahme/Ort des Geschehens
- Beschwerdebild mitsamt Details und Dauer/Verlauf
- Patienten- und Familienanamnese
- Ergebnis der physikalischen Untersuchung
- Optional erste Untersuchungsergebnisse/Befunde
- Optional erste therapeutische Schritte

Weitere wichtige Aspekte sind – wenn auch teilweise wiederholend – der Fokus auf wichtige Konzepte, das Vorhandensein aller Informationselemente, die für die Lösung notwendig sind, das Vermeiden von unnötigen Verkomplizierungen und negativen Formulierungen.

Verfeinernd weisen Case und Swanson auch auf den Bezug zum realen medizinischen Alltag hin: Echte Patienten als Vorlage für die Fragen, realistisches Befundmaterial und auch die echte Sprache der Patienten, die auch in direkter Rede wiedergegeben werden kann.

In einer anderen Arbeit werden ähnliche Empfehlungen abgegeben; in dieser wird abschließend auch noch einmal auf die Wichtigkeit des Review-Prozesses hingewiesen (Schuwirth 1999):

- Fälle sollen sich auf reale Patienten beziehen
- Beschreibung so eindeutig und klar wie möglich
- Ausreichend klinische und kontextuelle Informationen
- Ausreichend negative Informationen
- Keine vorinterpretierten Informationen (sondern *roh*)
- Essentielle Probleme mit klaren Fragestellungen
- Richtige und falsche Optionen eindeutig zu verteidigen

„Verbosity, window dressing, and red herrings“, lautet der Anfang des Titels einer wissenschaftlichen Arbeit von 1996 (Case et al. 1996). Dabei werden das Ausschmücken und umfassende Gestalten eines Items, sowie das teilweise Versehen mit falschen Hinweisen beleuchtet. Auch wenn es keine eindeutige Antwort gibt, ob keine, eine kurze oder eine lange Vignette besser sind, sind Case und Swanson überzeugt, dass Vignetten das Anwenden von Wissen prüfen. Darüber hinaus machen rohe, also nicht interpretierte Informationen in der Vignette eine Frage schwerer.

Genau diese Sichtweise deckt sich auch mit der an der Medizinischen Universität Graz. Smolle nennt in seinem Leitfaden zur Fragenerstellung, neben Authentizität und Kontextbezogenheit auch noch Situiertheit und Immersivität. Das bedeutet, dass die Fragen räumlich und zeitlich realistisch eingebettet sein sollen und die Studierenden sich gut in die geschilderten Situationen hineinversetzen können müssen – sie sollen *in ihnen aufgehen* (Smolle 2008).

C.F. Herreid zitierend, ergänzt er folgende Punkte: Man soll eine Geschichte erzählen und dabei auf einen interessanten Aspekt fokussieren, der in der Gegenwart oder nahen Vergangenheit angesiedelt ist. Man soll Empathie zu den geschilderten Personen herstellen, die direkte Anrede verwenden – diese Beschreibung endet dann in einem Konflikt, der vom Studierenden eine Entscheidung verlangt.

Er beschreibt auch eine Methode, um den Schwierigkeitsgrad einer Frage verändern zu können, und nennt diese dabei *Typikalität*. Auch wenn es bei der Beschreibung von Patientenfällen als Problemstellung häufiger Anwendung findet, so ist das Beschreiben eines mehr oder weniger *typischen* Sachverhaltes in beinahe allen fachlichen Bereichen möglich.

4.3. QUALITÄTSSICHERNDE MAßNAHMEN

Beispiele, wie sie Weih et al. in ihrer Publikation 2009 aufzeigen (Weih et al. 2009), unterstützen die Annahme, dass sich der Aufwand, den gesamten Prozess der Fragenerstellung inklusive Review sowie die Auswertung zu optimieren, auszahlt. Er beschreibt die Abschaffung problematischer Fragetypen, die Einführung valider Fragenformate sowie die Verlängerung der Fragenstämme, aber auch die Einführung eines Peer-Reviews. Damit kann er eine Zunahme des Schweregrades um 18 % und der Trennschärfe um 67 % zeigen. Rotthoff hat 2006 (Rotthoff & Soboll 2006) eine ähnliche Qualitätsverbesserung nach der Einführung von Workshops und einem Review-Komitee beschreiben können.

Ähnliche Bilder zeigen sich auch, wenn man sich die Fragen am Ende von CME-Fortbildungen in Journalen ansieht (Öchsner & Böckers 2013): Weniger als 20 % der Fragen haben einen Bezug zum Problemlösen und können als anwendungsorientiert gesehen werden und nur die Hälfte der Fragenitems sind ohne Lösungshinweise oder Fehler.

Dass sich die Interventionen auch lohnen, zeigt – neben den oben aufgezählten Beispielen – unter anderem der Scientific Report von Abdulghani (Abdulghani et al. 2015): Nach einem langfristig angelegten Fortbildungsprogramm für neue Lehrpersonen können zahlreiche Kennzahlen verbessert werden. Die Item-Schwierigkeit und die Trennschärfe der neuen Fragen verbessern sich, der kognitive Level nach Bloom ist höher und es werden weniger nicht-funktionierende Distraktoren und Lösungshinweise beschrieben.

Grob zusammengefasst hat man unter anderem folgende Möglichkeiten, die Fragen- und damit die Prüfungsqualität von MC-Prüfungen zu verbessern:

1. Schulung aller beteiligten Personen: Es sollen alle Personen geschult werden, die Fragenitems erstellen und schreiben, ebenso wie die Personen, die diese Items begutachten und mit diesen im Sinne der Prüfungszusammenstellung arbeiten. Diese Personen sollen die wesentlichen und erforderlichen Grundlagen der Lehre und des Prüfens vermittelt bekommen, ebenso wie die wichtigsten Empfehlungen, Regeln und Kriterien für die Erstellung hochwertiger Fragen.
2. Peer-Review: Alle Fragen, die in high-stakes-Prüfungen zu Einsatz kommen, sollen sorgfältig von mehreren instruierten Personen gegengelesen und begutachtet werden. Details dazu befinden sich im Unterkapitel.
3. Analyse der Prüfung, aller Items und der Antwortoptionen im Anschluss an eine Prüfung: Die Details dazu befinden sich im nachfolgenden Abschnitt Item- und Distraktorenanalyse.

4.3.1. SCHULUNGSMAßNAHMEN

Wie schwierig es sein kann, hochwertige MC-Fragen zu erstellen, zeigt eine Arbeit mit analysierten Prüfungen, in denen 28-75% der Fragen mangelhaft waren (Tarrant & Ware 2008). Auf der anderen Seite können sowohl das notwendige Wissen wie auch die notwendige Erfahrung erlernt und trainiert werden.

Den Bildungseinrichtungen muss die Bedeutung des Prüfens und der damit verbundenen Notwendigkeiten bewusst sein, um die richtigen Vorbereitungen treffen zu können.

Grundsätzlich sind mehrere Komponenten bei der Erstellung hochwertiger Fragen entscheidend. Die Grundvoraussetzung beim Fragenerstellen ist die Expertise im betroffenen Fachgebiet. Wer diese nicht besitzt, hat Schwierigkeiten, valide und saubere, eindeutige Fragenitems des passenden Schwierigkeitsgrades zu schreiben. Daneben sind die Grundkenntnisse des Prüfens und des Fragenschreibens entscheidend.

Vor allem die Empfehlungen aus der internationalen Literatur betreffend die Fragetypen, den Aufbau und die Formulierung sind hier zu vermitteln. Diese Kenntnisse vorausgesetzt, fällt abschließend dem Einüben und dem Training eine entscheidende Rolle zu. An vielen Stellen wird davon gesprochen, dass Fragenschreiben eine Kunst ist; es kann aber auch als Handwerk verstanden werden, bei dem die Fertigkeit mit häufigem Üben verbessert wird.

Die Notwendigkeit von Schulungsmaßnahmen soll hier erwähnt und betont werden. Details zum Umfang, zu den Inhalten und zur Tiefe, sowie zur Anrechnung, Zertifizierung und ähnlicher Aspekte sollen hier nicht ausführlich aufgeführt werden. Ergänzend und abschließend kann aber darauf hingewiesen werden, dass nicht nur das Schulungsangebot entscheidend ist, sondern die Autoren und Autorinnen auch auf andere Weise unterstützt werden können. So ist das Bereitstellen von Literatur, Empfehlungen und Checklisten, aber auch Lernzielkatalogen und Benutzerhandbüchern eingesetzter Software hilfreich und unterstützend.

4.3.2. PEER-REVIEW-PROZESS

Unter Peer-Review versteht man das Ansehen und Beurteilen einer Arbeit, in diesem Fall einer MC-Frage, durch *Peers*, also Kolleginnen und Kollegen. Wenn auch nicht zwingend, so erfolgt der Review zumeist vor dem Einsatz bei einer Prüfung. Die Personen, die den Review oder die Begutachtung durchführen, sollen die Frage lesen, beurteilen und bei Notwendigkeit kommentieren.

Die Beurteilung kann anhand einer Liste von Kriterien erfolgen, der Kommentar erfolgt in den meisten Fällen als Freitext. Als Beispiel wurde hier in Abbildung 11 das Reviewfenster des Item Management Systems (IMS) eingefügt (auf das System selbst wird im Kapitel 4.6 näher eingegangen).

Anhand dieses ist auch die Trennung von inhaltlichen und formalen Kriterien ersichtlich: Formale Kriterien zum Fragentyp, zu Lösungshinweisen, Beantwortbarkeit und Homogenität der Antwortoptionen, inhaltliche Kriterien zum Lernziel, zum Inhalt, dem Niveau, dem Anwendungsbezug und zu Verständlichkeit und Plausibilität, um nur ein konkretes Beispiel aus einer aktuellen Software zu bringen.

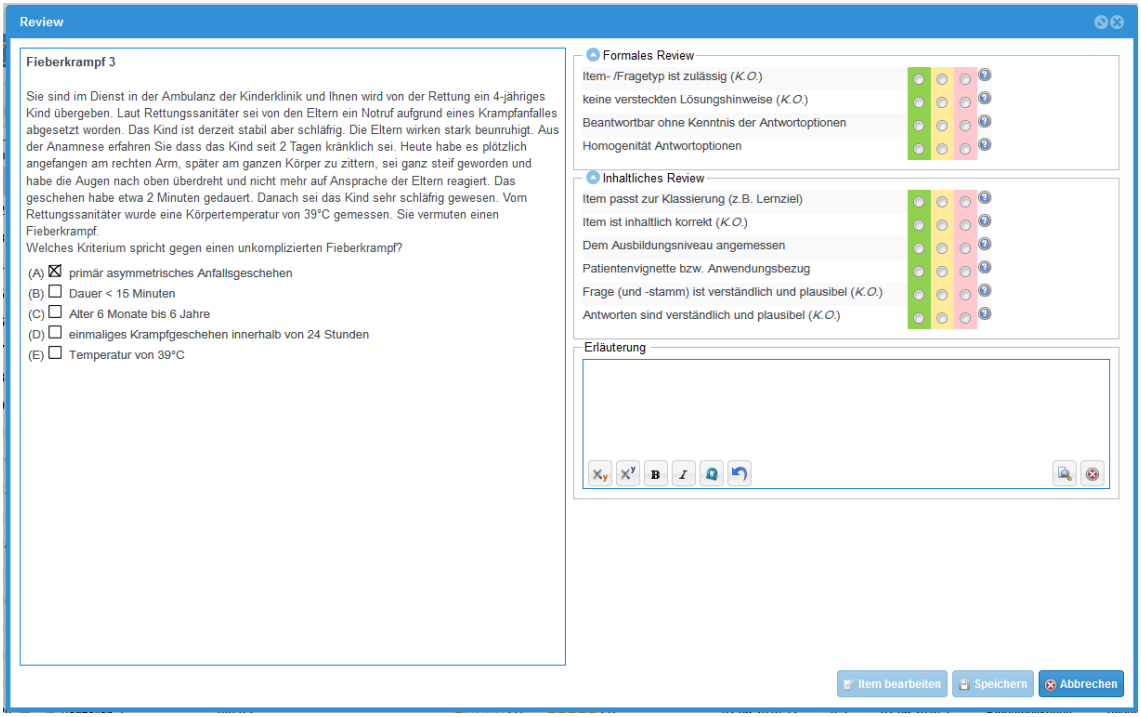

Abbildung 11: Screenshot des Review-Fensters in der Verwaltungssoftware IMS. Links die MC-Frage in voller Länge, rechts die Review-Maske und Freitext-Feld.

Es gibt noch zahlreiche Aspekte, den Review-Prozess betreffend. So muss zu Beginn geklärt werden, wie viele Begutachtungen notwendig sind, wer diese durchführt und wie man auf Kritikpunkte oder Verbesserungsvorschläge reagieren soll oder muss. Gerade bei der Frage, wer die Begutachtung durchführen soll, kann man anführen, dass sowohl fachlich kompetente Personen, wie auch fachfremde Personen geeignet sind. Fachfremde Personen haben einen neutraleren Standpunkt, haben möglicherweise eine breitere Sichtweise und sind ebenso gut geeignet, Rechtschreibfehler, Tippfehler und andere formale Probleme zu erkennen und auf diese aufmerksam zu machen.

Daneben hat man, auch in der Software IMS, die Möglichkeit eines Einzel- oder eines Gruppenreviews. Auch wenn der Zeitaufwand für Gruppenreviews etwas höher liegt, spricht beispielsweise Böhme von der Universität Freiburg von einer leichten Präferenz von Gruppenreviews (Böhme et al. 2012).

4.4. PRÜFUNGSAUSWERTUNG

Der vorliegende Abschnitt Prüfungsauswertung widmet sich der Auswertung im Sinne der Beurteilung und der damit verbundenen Notenfindung. Analysen als Basis von qualitätsverbessernden Prozessen werden im darauffolgenden Abschnitt behandelt.

Das Schema „ein Punkt bei korrekter Beantwortung einer MC-Frage und kein Punkt bei falscher Beantwortung“ werden hier vorausgesetzt. Je mehr Punkte ein Kandidat oder eine Kandidatin bei der Prüfung erreicht, desto besser das Ergebnis und desto besser die Note. Bei der Standardsetzung wird zu Beginn jene Mindestpunkteanzahl festgelegt, die für eine positive Endbeurteilung notwendig ist.

Die Person, die mit der Auswertung betraut ist, sollte diesen Prozess systematisch angehen. Zu Beginn sollte man sich das Gesamtergebnis ansehen und die Gesamtpunkte aller Studierenden auflisten oder als Histogramm darstellen. Damit kann man sehr schnell Asymmetrien und Auffälligkeiten wie auch Ausreißer identifizieren. Eine auffällig niedrige Punktezahl könnte beispielsweise auf eine abgebrochene Prüfung hinweisen.

In weiterer Folge werden die Prüfungen aller Studierenden ausgewertet und überprüft. Am Ende werden die Ergebnisse mitsamt den Noten von den Fachpersonen gegengeprüft und freigegeben. Oft kommt es im Anschluss zu Prüfungseinsichtnahmen und auch nachträglichen Streichungen von Fragenitems, was eine Nachberechnung aller Prüfungen notwendig macht. Allerdings werden diese Prozesse und Teilschritte in der jeweiligen Einrichtung sehr individuell gehandhabt, deshalb soll hier auch nicht näher darauf eingegangen werden.

4.5. ITEM- UND DISTRAKTORENANALYSE

Die nachfolgenden Aspekte der Itemanalyse basieren auf der Arbeit von Möltner, Schellberg und Jünger. Sie haben die Einzelaspekte im notwendigen Umfang gut und übersichtlich dargestellt (Möltner et al. 2006).

4.5.1. ITEMANALYSE

Die Berechnung der *Aufgaben-Schwierigkeit* (auch *Item-Schwierigkeit* oder englisch als *item difficulty* bezeichnet) ist definiert als die mittlere bei dieser Aufgabe erreichte Punktezahl und liegt zwischen 0 und 1, damit sind leichte Aufgaben durch hohe Werte charakterisiert und umgekehrt. Beachten muss man die Tatsache, dass dieser Kennwert immer in Bezug zur Prüfungspopulation gesehen werden muss, sie ist keine Eigenschaft einer bestimmten Aufgabe selbst. Als Richtwert kann ein Bereich zwischen 0,4 und 0,8 gesehen werden. Die Asymmetrie ist durch den Umstand bedingt, dass zu viele schwere Aufgaben die Motivation beeinträchtigen könnten. Zu betonen ist auch, dass jede Prüfung mit ihren Aufgaben den gesamten Bereich abdecken soll, da sowohl im oberen als auch im unteren Leistungsspektrum eine Differenzierung der Studierenden notwendig ist. Der Großteil der Aufgaben soll sich im mittleren Schwierigkeitsbereich befinden, hier sollte die Differenzierung des Großteils der Studierenden erfolgen. Abschließend soll auch auf die Faktoren Zeitmangel und Zufallstreffer hingewiesen werden. Beide können die Aufgabenschwierigkeit verzerren und damit eine sinnvolle Interpretation erschweren.

Die Berechnung der Aufgaben-*Trennschärfe* (auch *Item-Trennschärfe oder englisch als discrimination index bezeichnet*) dient der Überprüfung, ob die Aufgabe zwischen *leistungsstarken* und leistungsschwachen Studierenden zu unterscheiden vermag. Sowohl der *Diskriminationsindex D*, wie auch der häufiger verwendete *Korrelationskoeffizient r* nach Pearson-Bravais können dafür herangezogen werden.

Der *Diskriminationsindex D* ist die Differenz der mittleren Schwierigkeit der stärksten und der schwächsten Studierenden, wobei unterschiedlich sowohl die oberen und unteren 27 %, oder auch die oberen und unteren 33 % herangezogen werden.

Der *Korrelationskoeffizient r* gibt die Korrelation zwischen den erreichten Punktzahlen bei der untersuchten Aufgabe und den Punktzahlen der gesamten Prüfung wieder. Grob könnte man festhalten, dass D die Größe des Unterschieds beschreibt, r eher als Schärfe der Auftrennung gesehen werden kann. Trennschärfen von über 0,3 gelten dabei als gut, zwischen 0,2 und 0,3 als akzeptabel, zwischen 0,1 und 0,2 als marginal und unter 0,1 als schlecht. Negative Werte sind sehr schlecht, da sie einer vernünftigen Trennung zwischen guter und schlechter Gesamtleistung entgegenwirken.

4.5.2. DISTRAKTORENANALYSE

Sowohl die Item-Schwierigkeit wie auch die Item-Trennschärfe sind Kennwerte auf Itemebene. Mitunter sehr interessant und wichtig sind aber auch Analysen auf Antwortenebene. Hier sieht man sich alle Antwortoptionen und damit alle Distraktoren an, daher auch die Bezeichnung. Sie hat zum Ziel, zu leichte, zu schwere Antwortoptionen oder eine schlechte Abgrenzung der Distraktoren und nicht eindeutige Formulierungen aufzudecken. Auf der anderen Seite können mit dieser Analyse auch häufige und typische Studierendenfehler aufgedeckt werden. Entweder werden dabei die relativen Häufigkeiten der Antworten oder die relativen Häufigkeiten der falschen Antworten bezogen auf alle falschen Antworten angegeben.

Bei einer guten Frage (wie beispielsweise im linken Bild der Abbildung 12) soll die korrekte Option häufiger als jede inkorrekte Option gewählt worden sein, die Häufigkeiten der inkorrekten Optionen sollen gleichmäßig verteilt sein. Eine Antwortoption, die, eine entsprechend hohe Teilnehmerzahl bei der Prüfung vorausgesetzt, sehr selten oder nie gewählt wurde, ist zu überdenken. In der rechten Abbildung 12 wurden beispielsweise Antwortoption C und D nie gewählt, hier ist die Sinnhaftigkeit oder Plausibilität in Frage zu stellen. Ebenso auffällig und negativ zu werten ist der Umstand, dass die korrekte Option B von weniger Studierenden gewählt wurde, als der Distraktor E. Das kann auf eine mehrdeutige Formulierung oder Unzulänglichkeiten in der Lehre hinweisen.

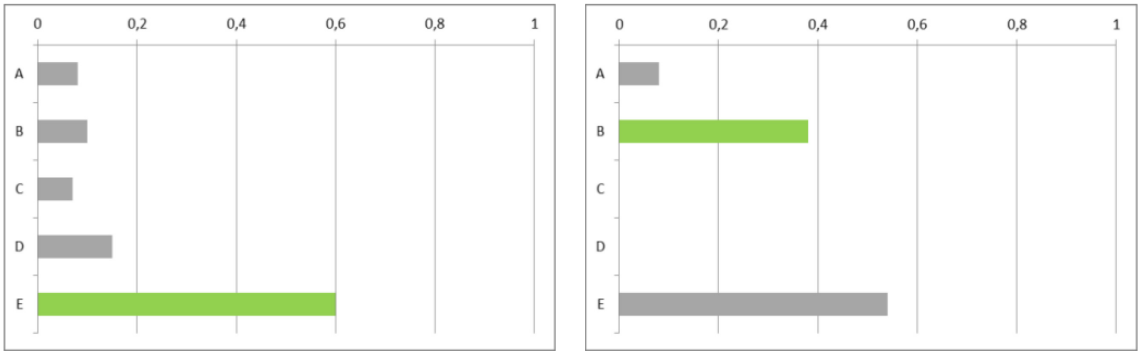

Abbildung 12: Zwei Beispiele, wie eine Distraktorenanalyse grafisch aussehen könnte. Neben den Antwortoptionen A bis E sind jeweils die relativen Häufigkeiten, mit der die Antworten gewählt wurden, als Balken dargestellt. Grün ist die korrekte, zu wählende Antwort, grau die Distraktoren.

Bei der Analyse der Antwortoptionen ist die Berücksichtigung der Trennschärfe der Antwortoptionen möglich, soll hier jedoch nur erwähnt, jedoch nicht ausgeführt werden.

4.5.3. Weitere Aspekte

Möltner kommt in seiner Arbeit im Anschluss auf die Reliabilität zu sprechen, die hier bereits in einem der früheren Kapitel angesprochen wurde (Möltner et al. 2006).

Er verweist auf die Formen der Reliabilität sowie auf die interne Konsistenz und den Koeffizienten Cronbachs α. Dazu passend behandelt er auch die wichtigsten Aspekte der Validität, darunter auch die Konstrukt-Unterrepräsentation und die konstruktirrelevante Varianz, die beide zu vermeiden sind. 2010 hat eine Schweizer Expertengruppe rund um Kropf die Auswirkungen von Itemanalyse-Besprechungen untersucht und publiziert (Kropf et al. 2010). Sie haben beschrieben, dass die Fragen nach Einführung dieser Besprechungen relevanter waren, deutlich trennschärfer waren und die Prüfungen insgesamt zuverlässiger gemessen haben.

4.6. Die Situation an der Medizinischen Universität Graz

Das Studium der Humanmedizin ist an der Medizinischen Universität Graz ein Diplomstudium und bereitet die Studierenden auf den zukünftigen Beruf als Arzt/Ärztin vor. Theoretische Grundlagen und praktische Fertigkeiten werden in einer integrativen und themenzentrierten Weise vermittelt[13].

Das Studium ist in drei Abschnitte gegliedert, wobei der erste Abschnitt zwei Semester, der zweite Abschnitt acht Semester und der dritte Abschnitt wiederum zwei Semester dauert. Der Gesamtumfang der Studienzeit beträgt 360 ECTS-Punkte oder 60 ECTS-Punkte pro Studienjahr[14].

Alle grundlegenden Informationen zum Studium der Humanmedizin und spezielle Ausschnitte wurden dem Studienplan der Version 11a[15] entnommen, welcher mit 1. Oktober 2014 in Kraft getreten ist. Ab der Version 10[16] aus dem Jahre 2012 wurden nur Änderungen bezüglich des Klinisch-praktischen Jahres durchgeführt, jedoch keine für diese Arbeit relevanten Abschnitte verändert.

[13] Offizielle Formulierung zum Studium der Humanmedizin (http://www.medunigraz.at/themen-studieren/humanmedizin/), abgerufen am 1. Juli 2015.
[14] ECTS-Punkte sind die Einheiten des ECTS-Systems, welche im Abkürzungsverzeichnis kurz erklärt wird.
[15] Aktueller Studienplan (http://www.medunigraz.at/themen-studieren/humanmedizin/studienplan/), abgerufen am 1. Juli 2015.
[16] Studienplanarchiv (http://www.medunigraz.at/themen-studieren/humanmedizin/studienplan/studienplanarchiv/), abgerufen am 1. Juli 2015.

Aktuell (1. Juli 2015) gibt es zusätzlich eine Version 13, die der Version 12 gefolgt ist, beide beziehen sich auf das weiterentwickelte Studium, welches ab dem Herbst 2014 sukzessive eingeführt wird, welches für diese Arbeit aber nicht relevant ist.

4.6.1. CURRICULUM UND PRÜFUNGSARTEN

Das Curriculum des Studiums der Humanmedizin ist ein Modul- und Track-System, der Unterricht findet fächerübergreifend und themenzentriert statt. Die Module sind Blockveranstaltungen zu je fünf Wochen[17]. Daneben gibt es noch mitlaufende Tracks, sowie Spezielle Studienmodule und Wahlfächer, die aber nicht Gegenstand der vorliegenden Arbeit sind.

Der Kern des Curriculums besteht aus 25 Modulen aus fünf Studienjahren. Die Module sind von M01 bis M29 nummeriert, lediglich M09, M18, M24 und M27 existieren aus entwicklungstechnischen Gründen aktuell nicht. Das sechste Studienjahr heißt auch Klinisch-praktisches Jahr und ist dem Sammeln klinischer Erfahrung gewidmet.

Laut dem oben erwähnten Studienplan in der Version 11a werden die Prüfungen so gestaltet, dass sie nachvollziehbar, reliabel, valide und somit für die Überprüfung der verschiedenen Lernziele – Wissen, Fertigkeiten und Einstellungen – geeignet sind[18]. Neben Fachprüfungen (schriftlich oder mündlich) gibt es auch Lehrveranstaltungen mit immanentem Prüfungscharakter (PR, Se, Ue, SU und Ex), mündlich kommissionelle Gesamtprüfungen (OSKE) und Formative Prüfungen (PTM).

Es gibt 25 Fachprüfungen, auch Modul-Abschlussprüfungen genannt, ohne auf andere Curriculums-Bestandteile wie Tracks, OSKE-Prüfungen oder PTM-Teilnahmen einzugehen. Bei den Fachprüfungen werden unterschiedliche Prüfungsformate genutzt, allen voran schriftliche und mündliche Formate, wobei bei

[17] Formulierung zum Aufbau des Studiums (http://www.medunigraz.at/themen-studieren/humanmedizin/pflichtmodule-und-tracks/), abgerufen am 1. Juli 2015.
[18] Formulierungen zu den Prüfungen dem Studienplan in der Version 11a entnommen.

den schriftlichen Formaten sowohl offene, wie auch geschlossene Fragenformulierungen genutzt werden.

Prüfungsformat		Module
Schriftliches Format		
	Geschlossener Fragentyp (Multiple-Choice)	M01-M03, M04*, M05, M06-M08, M10-M12*, M14, M16, M17, M19, M21-M23, M25-M29
	Offener Fragentyp (Short-Answer/Short-Essay)	M10-M12*, M13, M15, M20
Mündliches Format		M04*

Tabelle 1: Auflistung der Prüfungsformate, ergänzt um die Module, die diese anwenden (mit * sind Module gekennzeichnet, die mehr als ein Prüfungsformat verwenden).

Von den 25 Fachprüfungen nutzen 22 zur Gänze oder zum Teil das MC-Format (88 %), lediglich drei Module prüfen ausschließlich mit dem Short-Answer-Format. Diese Module und jene, die Prüfungsformate kombiniert einsetzen, sind in Tabelle 1 aufgelistet.

Im Folgenden werden kurz die drei Phasen Vorbereitung, Durchführung und Nachbereitung oder Auswertung einer Prüfung erläutert: Die Lehrenden der Kliniken und Institute der Medizinischen Universität Graz sind angehalten, Prüfungsfragen zu ihrem Fachgebiet zu schreiben.

Die erstellten Fragenitems sollen im Anschluss entweder von Fachkolleginnen und Fachkollegen oder von Prüfungsexpertinnen und Prüfungsexperten begutachtet werden, was zu einem mehrstufigen Überarbeitungsprozess führen kann. Diese Schritte, der Erstellung, Begutachtung und Speicherung werden in der Verwaltungssoftware IMS durchgeführt.

Anschließend werden die Prüfungsfragen von den Prüfungskoordinatorinnen und Prüfungskoordinatoren (Lehrende mit der entsprechenden Funktion) unter bestimmten Gesichtspunkten ausgewählt und zusammengestellt. An dieser Stelle muss erklärt werden, dass es an der Medizinischen Universität Graz zwei Arten der Durchführung einer schriftlichen MC-Prüfung gibt.

Auf der einen Seite gibt es die Durchführung als Papier-basierte Prüfung (auch als Paper-Pencil bezeichnet), diese wird von 9 Modulen (M01-M03, M05, M14, M17, M21-M23) genutzt, sowie von weiteren 4 Modulen in Kombination mit anderen Formaten (M04, M10-M12). Das ergibt in Summe 13 Module, die Papier-basiert prüfen. Diese Durchführung wird im folgenden Abschnitt erläutert, danach folgt der Abschnitt der Computer-basierten Prüfung.

4.6.2. PAPIER-BASIERTE PRÜFUNG

Bei Papier-basierten Prüfungen wird die Durchführung zusammen mit der Prüfungsabteilung vorbereitet, die Durchführung selbst (im reservierten Prüfungsraum) obliegt den Prüfungsverantwortlichen. Die Prüfung besteht in der Regel aus 60 ausgewählten Fragenitems, die die Prüfungskoordinatorinnen und Koordinatoren zusammenstellen. Die Prüfungszeit beträgt dabei in der Regel 75 Minuten und die Bestehensgrenze liegt normalerweise bei 66 % korrekt beantworteter Fragen. Lediglich bei Modul 14 beträgt die Zeit 90 Minuten. Die Prüfung wird zusammengestellt und entsprechend der Teilnehmerzahl ausgedruckt. Mehrere unterschiedliche Versionen/Gruppen zu erstellen ist möglich, dabei wird nur die Item-Reihenfolge verändert. Die Prüfung wird in einem entsprechend großen Prüfungsraum abgehalten, die Fragebögen werden zusammen mit den Antwortbögen ausgeteilt. Auf Ersterem stehen die 60 Fragenitems, auf dem Antwortbogen werden die korrekten Antworten von den Studierenden mit einem Bleistift markiert.

Die Antwortbögen werden im Anschluss an die Prüfungsabteilung übermittelt, welche diese dann einscannt und automatisiert auswertet. Die Ergebnisse werden an die koordinierenden Personen weitergeleitet; erst nach deren Bestätigung kann die Beurteilung offiziell freigegeben werden.

4.6.3. COMPUTER-BASIERTE PRÜFUNG

Wie bereits erwähnt, gibt es auf der anderen Seite die Durchführung als Computerprüfung (auch als Online-Prüfung bezeichnet), diese wird von 9 Modulen genutzt (M06-M08, M16, M19, M25-M29).

Diese Prüfung besteht in der Regel ebenfalls aus 60 ausgewählten Fragenitems, die Prüfungszeit beträgt dabei in der Regel 60 Minuten – zum Teil auch 75 Minuten – und die Bestehensgrenze liegt normalerweise bei 66 % korrekt beantworteter Fragen. Der Unterschied zur Papier-Variante liegt in der Durchführung der Prüfung. Die Studierenden buchen sich einen Computer-Prüfungsplatz für einen bestimmten Zeitslot. In dieser reservierten Zeit sitzen Sie vor dem Computer in einem überwachten Prüfungsraum und absolvieren ihre gewünschte Prüfung. Die Fragenitems werden per Zufall vom Computer gezogen, v.a. um zu verhindern, dass mehrere Studierende dieselbe Prüfung absolvieren und sich zwischenzeitlich absprechen können. Damit verbunden muss der Fragenpool größer sein.

Bei Computer-basierten Prüfungen werden ein vorläufiges Ergebnis und eine vorläufige Note sofort berechnet. Am Ende des 2-Wochen Zeitraumes, in dem Online-Prüfungen abgelegt werden können, werden die Prüfungen an die Prüfungsabteilung übermittelt, von dieser geprüft und anschließend der koordinierenden Person zur Bestätigung weitergeleitet. Erst nach dieser kann die Prüfungsabteilung die Beurteilung endgültig freigeben.

Nachteilig ist dabei der Zufall zusehen, da die Verantwortlichen auf Lehrenden-Seite nicht unmittelbar in die Inhalte-Verteilung und -Gewichtung eingreifen können. Zudem ist auch eine Qualitätskontrolle im Sinne einer Item- oder Distraktorenanalyse nicht möglich. Theoretisch ist anzumerken, dass eine Prüfung zu einem singulären Zeitpunkt möglich wäre, wenn der Computerprüfungsraum entsprechend groß wäre.

4.6.4. INHALTE UND WEITERBILDUNG

Wie bereits gesagt, sind die Lehrenden der Kliniken und Institute der Medizinischen Universität Graz angehalten, Prüfungsfragen zu ihrem Fachgebiet zu schreiben. Meistens wird dabei das Fachgebiet unter den Lehrenden aufgeteilt, sodass jeder Fachexperte/jede Fachexpertin Prüfungsfragen zu seinem/ihrem Spezialgebiet schreibt. Die Inhalte sollen sich im Weitesten an den Ausbildungszielen orientieren, im Konkreten an den zuvor definierten Lernzielen.

Dazu dient der 2013 fertig gestellte und veröffentlichte Klinische Lernzielkatalog der Medizinischen Universität Graz[19], der für den Zeitraum zwischen dem dritten und fünften Studienjahr gedacht ist. Er ist fachspezifisch gegliedert und mit Ergänzungen zur Lerntiefe versehen. Beim Fragenerstellen sollen sich alle Lehrenden an diesem orientieren. Ein Lernzielkatalog für den Vorklinischen Bereich ist zurzeit in Arbeit, er wird in einem mehrstufigen Prozess von zahlreichen wissenschaftlichen und nicht-wissenschaftlichen Mitarbeiterinnen und Mitarbeitern erstellt.

Inhaltlich werden neben Faktenfragen auch Verständnisfragen und besonders Anwendungsfragen gefordert. Hier sollte man sich an der – im Einführungsteil erwähnten – Lernzieltaxonomie von Benjamin Bloom orientieren. Alle erstellten Fragen sollen zudem dem empfohlenen Fragentyp-A positiv entsprechen. Daneben sollen negative Fragenformulierungen sehr sparsam verwendet werden, Fragen des Typs-K vollständig vermieden werden.

Eine Besonderheit, nämlich die Generierung neuer Fragen mittels Permutation[20], soll ebenfalls vermieden werden.

[19] Roller R., Dimai H.P., 2013. Klinischer Lernzielkatalog der Medizinischen Universität Graz. Wien: Manz Verlag. ISBN: 978-3-200-03422-8.

[20] In Graz wird unter Fragen-Permutation die Generierung von neuen Fragen mittels automatischer Zusammenstellung auf Basis von zahlreichen positiven und negativen Aussagen zum Fragenthema verstanden. Anders formuliert, wenn eine Liste positiv formulierter Aussagen und eine Liste negativer Aussagen vorhanden ist, kann man das zufällige Zusammenführen von einer positiven mit vier negativen Optionen (Typ-A positiv) oder einer negativen mit vier positiven Optionen (Typ-A negativ) automatisieren.

Sowohl die Details die formalen Kriterien betreffend, wie auch die inhaltlichen Aspekte werden in Schulungen im Rahmen der Internen Weiterbildung vermittelt. Im Rahmen der Internen Weiterbildung werden regelmäßig und mehrmals im Semester Schulungstermine zur Erstellung hochwertiger MC-Fragen angeboten. Einerseits gibt es Schulungstermine mit Vortragscharakter, die zwischen zwei und drei Stunden dauern und für Lehrende mit Zeitmangel gedacht sind. Andererseits gibt es ganztägige Termine, die als Workshops konzipiert sind, sowie Möglichkeiten individueller Betreuung beim Thema der Fragen-Erstellung.

Die Schulungsinhalte sind grundsätzlich dieselben und gliedern sich grob in drei Bereiche: formale Kriterien für hochwertige MC-Fragen, inhaltliche Überlegungen und das Erstellen von Klinischen MC-Fragen. Ergänzende Inhalte sind der Umgang mit dem Lernzielkatalog, der Verwaltungssoftware IMS und Überlegungen zu qualitätssichernden Maßnahmen.

4.6.5. GENUTZTE SOFTWARE

Alle Fragenitems, die für schriftliche MC-Prüfungen – unabhängig ob die Durchführung Papier- oder Computer-basiert funktioniert – benötigt werden, werden ab 2012/2013 in der Verwaltungssoftware IMS verwaltet. Die Umstellung auf dieses System hat in der zweiten Jahreshälfte 2012 begonnen und war in der ersten Jahreshälfte 2013 für alle Papier-basiert prüfende Module größtenteils abgeschlossen. In diesem System werden alle MC-Fragen gespeichert, für Papier-basierte Prüfungen ausschließlich, für Online-Prüfungen zusammen mit dem älteren Prüfungssystem Questionmark Perception.

Im Weiteren können im IMS auch Fragenitems für mündliche und praktische Prüfungen abgelegt werden, was jedoch im Zusammenhang mit dieser Arbeit nicht relevant ist.

Im IMS können Fragenitems direkt erstellt oder alternativ über Schnittstellen aus anderen Systemen importiert werden. Eine systematische Trennung und Kategorisierung ist ebenso möglich, wie das Filtern von Fragenpools. Alle Fragenitems werden in einer tabellarischen Ansicht angezeigt (wie in Abbildung 13 ersichtlich) und können separat oder gemeinsam angewählt werden. Diese Items können mit Kennzahlen versehen, begutachtet und kommentiert werden. Die Itemverwendung wird ebenfalls registriert, wie alle anderen Änderungen und Freigaben durch den Autor oder die Autorin.

Die statistischen Kennzahlen Item-Schwierigkeit und Item-Trennschärfe[21] werden berechnet und in der tabellarischen Auflistung angezeigt, ebenso wie auf den Prüfungsberichten.

Auf die Medizinische Universität Graz bezogen, sollen alle Lehrenden mit diesem System vertraut sein und seine/ihre Fragenitems selbst eingeben und pflegen. Pro Modul können alle Fragenitems der Kollegeninnen und Kollegen eingesehen, begutachtet und kommentiert werden, eine direkte Änderung durch diese muss jedoch vom Erstellenden extra erlaubt werden.

Die Software IMS wurde vom 2005 gegründeten Prüfungsverbund Medizin entwickelt. Da die Idee der Kooperation im Prüfungsbereich von anderen erkannt und geschätzt wurde, wurde aus dem Prüfungsverbund die Dachorganisation UCAN mit mehreren Netzwerken und Kooperationen (Hochlehnert et al. 2012)[22].

[21] https://www.ucan-assess.org/cms/tools/test-statistical-analysis/ (abgerufen am 15. September 2015).
[22] UCAN/IMS: https://www.ucan-assess.org/cms/de/ueber-uns/ (abgerufen am 8. Juli 2015).

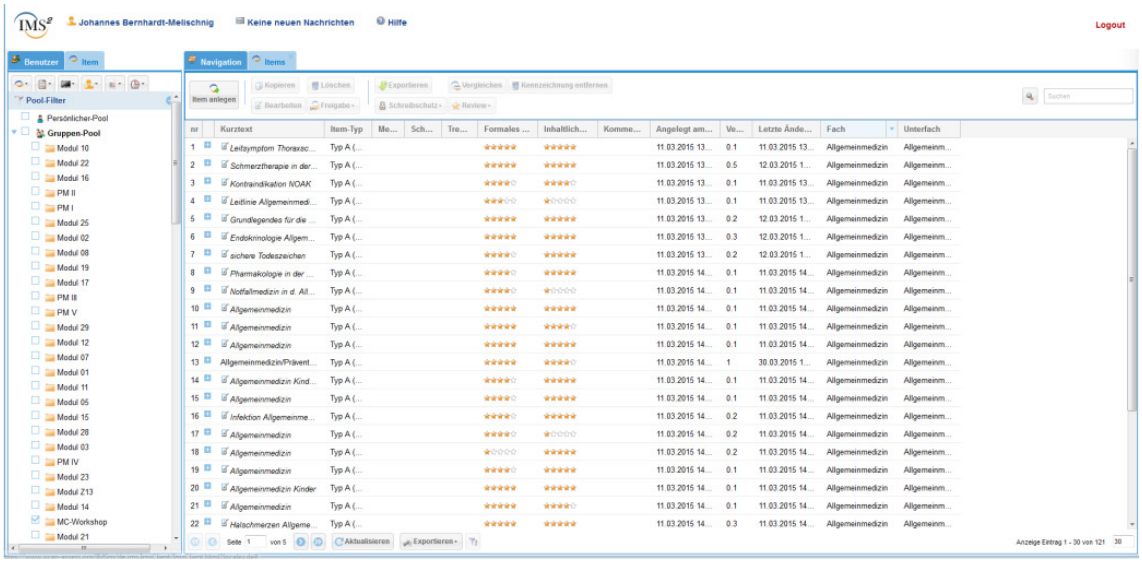

Abbildung 13: Screenshot der webbasierten Version der Software IMS, hier die Hauptansicht mit den Auswahloptionen links und der tabellarischen Auflistung der Items im großen rechten *Frame*.

Vor 2012 wurden die Fragenitems in einem System namens Questionmark Perception (QMP) verwaltet. Die Software Perception wurde von der Firma Questionmark entwickelt und würde aktuell in der Version 5.7 zur Verfügung stehen. Wegen einer speziellen Situation mit der Schnittstelle zu anderen Systemen ist an der Medizinischen Universität Graz nicht die aktuellste Version in Verwendung.

Eine Besonderheit an unserer Universität ist die Durchführung von Computer-Prüfungen. Bei denen werden die Fragenitems grundsätzlich in IMS erstellt, begutachtet und gespeichert, danach werden sie aber in QMP überführt, welches das verwendete System zur Durchführung von Computer-Prüfungen ist. Der Grund ist, das IMS die für die Grazer Medizinische Universität notwendige Online-Prüfungs-Funktionalität nicht zur Verfügung stellt. Dennoch ist IMS nicht zwingend notwendig, um das alte System QMP zu füllen, da dieses über diverse Schnittstellen verfügt, mit denen man Fragenitems importieren kann.

4.6.6. Auswertung und statistische Kennzahlen

Ein gravierender Unterschied liegt in der Auswertung und in der damit verbundenen Berechnung statistischer Kennzahlen. Bei einer Papier-basierten Prüfung bekommen alle Studierenden eines Prüfungsdurchganges dieselben Fragenitems, eventuell in unterschiedlicher Reihenfolge, aber dieselben. Wenn es sich um einen normalen Prüfungsdurchgang – also keinen Wiederholungstermin – handelt, beantworten zwischen 50 und 200 Studierende dieselben Fragenitems. Aus diesem Grund ist auch eine Berechnung der Kennzahlen wie beispielsweise Item-Schwierigkeit und Item-Trennschärfe legitim. Bei Computer-Prüfungen bekommen alle Studierenden höchst unterschiedliche Prüfungen angezeigt, da der Computer per Zufall aus einem mehr oder weniger großen Fragenpool zieht. Daher beantworten mitunter nur sehr wenige Studierende ein und dasselbe Fragenitem. Wenn weniger als 15 Studierende eine Frage beantworten, können nur bedingt aussagekräftige Kennzahlen berechnet werden. Die Berechnung ist theoretisch möglich, aber statistisch nicht belastbar.

4.6.7. Progress Test Medizin (PTM)

Der PTM ist ein formatives Prüfungsformat, welches von der Charité in Berlin durchgeführt wird. Die Bewertung dieser Prüfung hat keinen Einfluss auf den Studienfortschritt, die Grundidee ist, dass Studierende ihre Leistung mit den durchschnittlichen Leistungen anderer Studierenden vergleichen können.

Dabei werden Leistungen von Teilnehmenden gleicher Semester verglichen. Die Universität kann durch diesen Leistungsvergleich Rückschlüsse auf die Ausbildung und mögliche Mängel ziehen, eingeteilt in Fächergruppen und Organsysteme.

Studierende an der Medizinischen Universität Graz mit einem Studienbeginn vor dem WS 2012/2013 müssen den PTM im Laufe ihres Studiums mindestens zweimal absolvieren, ab diesem Semester müssen sie den PTM am Anfang des zweiten und vierten Studienjahres und im Laufe des sechsten Studienjahres verpflichtend absolvieren[23].

Dieser Wissenstest besteht aus 200 MC-Fragen mit Einfachauswahl und die Teilnehmenden haben drei Stunden Zeit. Für richtig beantwortete Fragen erhält man einen Punkt, für falsche Beantwortungen einen Punkt Abzug, eine gesonderte Option „weiß nicht" entspricht keiner Punkteveränderung. Die Fragenitems aus dem Pool stammen zum überwiegenden Teil von der Charité Berlin und unterliegen einer strengen Qualitätskontrolle. Eine gesonderte Vorbereitung ist nicht notwendig oder erwünscht, da der Wissensstand zu diesem Zeitpunkt im Studium von Interesse ist und verglichen wird.

[23] Informationen zum PTM (http://www.medunigraz.at/themen-studieren/humanmedizin/ptm/), abgerufen am 1. Juli 2015.

5. Material und Methoden

5.1. Daten

5.1.1. Motivation für die Auswahl

Da die Qualität von MC-Fragenitems an unserer Universität untersucht werden soll, stand die tabellarische Auflistung ebendieser im Vordergrund. Die beiden Versionen der Fragen-Verwaltungssoftware – also QMP und IMS – enthielten alle Fragenitems, mit denen geprüft wurde und die in dieser Arbeit untersucht wurden. Da die Veränderung der Verwaltungssoftware anfänglich eine Rolle gespielt hat, wurden ein komplettes Studienjahr mit dem alten System (Studienjahr 2011/2012) und ein komplettes Studienjahr mit dem neuen System (Studienjahr 2013/2014) ausgewählt. Eine vollständige Dokumentation der Fragen mitsamt den Antwortoptionen, ergänzt um Informationen, wer diese Items zu welchem Zeitpunkt erstellt hat, war notwendig.

Da bei der Untersuchung statistische Kennwerte wie Item-Schwierigkeit und Item-Trennschärfe notwendig sind, wurden nur Papier-basierte Prüfungen aus diesen beiden Jahren genommen. Online-Prüfungen haben, wie bereits beschrieben, den Nachteil fehlender Kennzahlen. Da der Item-Pool sehr groß ist und die Zuordnung von Items zu einzelnen Studierenden per Zufall erfolgt, können keine belastbaren Kennzahlen berechnet werden.

5.1.2. BESCHREIBUNG DER AUSWAHL

22 Module prüfen schriftlich mit geschlossenen Fragen, also MC-Fragen, davon führen 13 Module diese Prüfungen mit Papier durch – auf diesen liegt der Fokus. Alle Module bieten ihre Prüfungen mehrmals im Semester und im Studienjahr an, hier wird von mehreren Durchgängen gesprochen. Alle Module (10 Module: M01-M05, M10-M12, M14, M17) bieten grundsätzlich drei Durchgänge pro Semester und damit 6 Durchgänge im Jahr an, die Module M21, M22 und M23 (3 Module) bieten 4 Durchgänge pro Semester und damit 8 Durchgänge im Jahr an[24].

In Summe sind das 168 potentielle Durchgänge, die untersucht werden hätten können. Da jedoch die beiden statistischen Kennzahlen Item-Schwierigkeit und Item-Trennschärfe benötigt werden, müssen kleine Durchgänge, bei denen wenige Studierende teilnehmen ausgeschlossen werden. Der Grund liegt in der bereits beschriebenen fehlenden Belastbarkeit der Kennwerte. Beispielsweise lässt sich die Schwierigkeit eines Items nur schwer feststellen, wenn nur drei Studierende bei einem Wiederholungstermin angetreten sind und dieses Item beantwortet haben. Als Grenzwert wurden 25 Studierende genommen, d.h. es werden nur Durchgänge in die Analyse genommen, bei denen mindestens 25 Studierende teilgenommen haben. Dies trifft auf 90 Prüfungsdurchgänge zu (46 Durchgänge aus dem Jahr 2011/2012 und 44 Durchgänge aus dem Jahr 2013/2014).

5.1.3. WICHTIGE BEARBEITUNGSSCHRITTE

Die Daten – also alle Fragenitems nach Prüfungsdurchgängen sortiert – wurden aus den beiden Verwaltungssystemen QMP und IMS entnommen. Bei den Daten aus QMP wurde eine Excel-Variante exportiert, eine Datei für jede Prüfung. Diese Dateien wurden umgeformt, der Übersicht wegen gekürzt und zusammenkopiert.

[24] Die Auflistung von administrativen Details zum Angebot an Prüfungsdurchgängen sind dem Zweck der Arbeit entsprechend nur überblicksmäßig erfasst.

Die Grunddaten aus IMS wurden ebenfalls als Excel-Datei exportiert, die vollständigen Fragentexte wurden jedoch gesondert als pdf-Datei exportiert. Die Grunddaten, also die Liste der Items, wurden parallel zu den Item-Listen aus QMP zusammenkopiert.

Die Sichtung und Bewertung einer jeden einzelnen Frage wurde händisch durchgeführt, lediglich bei den Daten aus QMP (wo Item-Liste und Fragentexte gemeinsam vorgelegen sind) waren automatisierte Plausibilitätskontrollen möglich. Über alle Bearbeitungsschritte hinweg wurde sichergestellt, dass eine jede Frage bis zum ersten Datenexport nachvollziehbar bleibt. Dies betrifft sowohl die Betitelung als auch die Nummerierung.

Die Kategorisierung der Fragen-Items wurde händisch durchgeführt, die Ergebnisse händisch in die Haupttabelle (zusammengeführte Item-Liste) eingetragen und kontrolliert. Die wichtigen Kennzahlen Item-Schwierigkeit und Item-Trennschärfe waren von Beginn an Teil der Item-Listen.

Alle Export- und Bearbeitungsschritte wurden von sorgfältigen Kontrollen begleitet. Nach allen Umformungsschritten wurden Kontrollsummen errechnet und verglichen, ebenso Plausibilitätskontrollen durchgeführt.

5.1.4. DATENBEREINIGUNGEN

Bei den exportierten Datensätzen aus QMP gab es teilweise Schwierigkeiten, da die Software zu dem Zeitpunkt des Exportes etwa zehn Jahre alt und nur mehr bedingt zeitgemäß war. Somit wurden Eintragungen mit fehlenden oder offensichtlich falschen Zahlenwerten gelöscht. Nur wenn eine sinnvolle Korrektur der Eintragungen gewährleistet werden konnte, wurde diese durchgeführt, ansonsten wurden die fehlerhaften Eintragungen gelöscht. Auch ein systematischer Fehler bei der Berechnung der Item-Schwierigkeit unter besonderen Bedingungen wurde entdeckt und führte zur Löschung einiger weniger Einträge.

All diese Veränderungen führten dazu, dass sich die Anzahl der Items von anfänglich 2350 um 84 auf 2266 reduziert hat.

Bei den exportierten Datensätzen aus IMS gab es keine Probleme wie bei QMP beschrieben, die Schwierigkeit lag hier in erster Linie bei den getrennten Exporten: Auflistung der Items einerseits und vollständige Fragentexte andererseits. Dies machte eine händische Übertragung der Kategorisierungen notwendig. Letztendlich wurden auch hier Eintragungen mit fehlenden oder falschen Werten gelöscht oder wenn möglich korrigiert. Aus unterschiedlichen Gründen wurden im IMS Kennzahlen öfters nicht berechnet, womit sich die höhere Zahl an gelöschten Eintragungen erklärt. Bei diesen Daten reduzierte sich die Anzahl der Items von anfänglich 2491 um 227 auf 2264.

5.1.5. BESCHREIBUNG DER FINALEN DATENSÄTZE

Der erste Datenblock, welcher aus Questionmark Perception exportiert wurde und auf das Studienjahr 2011/2012 reduziert war, hat folgende Eigenschaften: 20 Prüfungsdurchgänge aus dem Wintersemester 2011 und 26 Prüfungsdurchgänge aus dem Sommersemester 2012 ergeben 2266 Datensätze, also MC-Fragenitems mit vollständiger Beschreibung.

Der zweite Datenblock, welcher aus dem ItemManagementSystem exportiert wurde und auf das Studienjahr 2013/2014 reduziert war, hat folgende Eigenschaften: 17 Prüfungsdurchgänge aus dem Wintersemester 2013 und 27 Prüfungsdurchgänge aus dem Sommersemester 2014 ergeben 2264 Datensätze, also MC-Fragenitems mit vollständiger Beschreibung.

Beide Datenblöcke und damit alle Items in der tabellarischen Auflistung weisen folgende Kennzahlen auf:

- Eindeutige Identifikationsnummer (zur Rückverfolgung)
- Studienjahr, Semester, Prüfungsdurchgang und Modulzugehörigkeit
- Nummer und Titel des Fragen-Items
- Fragetypen und Item-Eigenschaften (binäre Variablen zur inhaltlichen Beschreibung):
    - Typ-A positiv/negativ
    - Typ-K
    - Aussagen
    - Vignette
    - Abbildung
    - Rechnung (nur im Studienjahr 2013/2014)
- Spezielle Beschreibung der Antwortoptionen/Distraktoren
    - „NFD-5%“, Nicht-funktionierende(r) Distraktor(en) (von weniger als 5 % der Studierenden gewählt).
    - „NFD-0%“, Nicht-funktionierende(r) Distraktor(en) (von keiner/keinem der Studierenden gewählt).
    - „DiDo“, Distraktor dominiert
- Item-Schwierigkeit und Item-Trennschärfe

Im Folgenden soll kurz die Bedeutung der einzelnen binären Variablen (Fragetypen und Item-Eigenschaften) und ordinalen Variablen (Beschreibung der Distraktoren) und deren Beurteilung erläutert werden:

Typ-A positiv/negativ bedeutet, dass entweder eine normale und positiv formulierte Frage oder eine negativ formulierte Frage vorliegt, bei der Negationen wie „kein/keine/keiner“, „nicht“ oder „außer“ vorkommen, die zumeist durch Fettdruck oder Unterstreichung hervorgehoben werden.

K-Typ bedeutet das Vorliegen einer Frage dieses definierten Typs, welcher in Kapitel 3.3 und in Abbildung 5 erklärt wurde.

Aussagen-basiert bedeutet, dass die Frage auf den Vergleich unterschiedlicher Aussagen zu einem Thema abzielt. In den meisten Fällen lautet der Fragenbeginn: „Welche der vorliegenden Aussagen zum Thema xy ist richtig?"

Vignette bedeutet, dass es sich um eine praxisnahe und anwendungsorientierte Frage handelt, bei der auf einen Patientenfall oder auf ein klinisches Szenario verwiesen wird[25].

Abbildung bedeutet das Vorhandensein einer Abbildung (Grafik oder Foto) in der Frage, *Rechnung* bedeutet das Vorhandensein einer Rechnung, bei der mindestens mathematische Grundrechnungsarten gefordert sind.

Die Abkürzung „*NFD-5%*" bedeutet ausgeschrieben: Nicht-funktionierende Distraktoren, bzw. Distraktoren, die von weniger als 5 % der Teilnehmenden gewählt wurden. Diese Zahl liegt zwischen 0 und 4, vorausgesetzt, dass 5 Antwortoptionen zur Auswahl stehen. Je größer dieser Kennwert, desto größer ist die Zahl schlechter und wenig plausibler Antwortoptionen oder Distraktoren. Parallel dazu bedeutet „*NFD-0%*": Nicht-funktionierende Distraktoren, bzw. Distraktoren, die von keiner/keinem der Teilnehmenden gewählt wurde. Auch diese Zahl liegt zwischen 0 und 4, fünf Antwortoptionen vorausgesetzt. Die Bedeutung ist deckungsgleich mit der oberen Kennzahl.

Die Abkürzung „*DiDo*" steht für „Distraktor dominiert" und bedeutet, dass eine falsche Antwortoption, also ein Distraktor, von den Teilnehmenden häufiger gewählt wurde, als die richte Antwortoption. Dieser Umstand kann auf eine missverständlich formulierte Frage hinweisen.

[25] Erläuterung: ein solcher Fall oder ein solches Szenario ist für die Lösung der Frage notwendig. Die Frage kann also ohne die Beschreibung zuvor nicht gelöst werden. Im Gegensatz dazu stehen Beschreibungen, die die Frage „ausschmücken", aber für die Beantwortung nur bedingt relevant sind.

5.1.6. DATEN DES PTM

Die Medizinische Universität Graz nimmt seit vielen Jahren am PTM teil. Eine Übersicht über das Gesamtergebnis in Berichtsform (pdf-Format) existiert seit dem Wintersemester 2010/2011, ein detaillierter Auswertebericht allerdings erst seit dem Sommersemester 2012. Dieser Umstand ist entscheidend, da in diesem Auswertebericht die durchschnittlich erreichten Punktewerte der jeweiligen Jahrgänge, sowohl der Medizinischen Universität Graz, wie auch der anderen Universitäten/Fakultäten (als Vergleichsgruppe) enthalten sind. Die untenstehende Abbildung 14 einer Tabelle aus dem Sommersemester 2014 zeigt die Mittelwerte, die Standardabweichungen und die Teilnehmerzahlen für beide Gruppen (Medizinische Universität Graz und andere Fakultäten) ebenso wie die Mittelwertsdifferenzen, die gepoolten Standardabweichungen und die Effektstärken.

Tabelle 6.1.: Globaler Vergleich

	Eigene Teilnehmer			Andere Fakultäten					
Semester	MW	SD	N	MW	SD	N	MWdiff	SDpooled	ES
1				9.91	14.54	572			
2				15.48	12.42	937			
3				20.68	14.25	495			
4	16.33	14.33	21	30.97	17.04	1665	-14.63	17	-0.86
5	24.85	21.06	13	38.68	21	592	-13.84	20.96	-0.66
6	35.14	18.24	86	44.83	21.98	1505	-9.69	21.79	-0.44
7	38.86	24.85	7	48	23.06	587	-9.14	23.04	-0.4
8	54.16	22.04	101	53.9	26.88	1069	0.25	26.48	0.01
9	42.79	15.97	14	59.62	25.65	521	-16.83	25.41	-0.66
10	58.87	22.2	71	68.32	28.66	1242	-9.45	28.33	-0.33
11	59.34	26.82	32	78.75	29.23	130	-19.41	28.6	-0.68

Abbildung 14: Screenshot der Tabelle zum globalen Vergleich aus dem Auswertebericht für die Medizinische Universität Graz aus dem Sommersemester 2014.

Für einen Überblick über die Leistung der Studierenden der Medizinischen Universität Graz wurden alle verfügbaren Semesterberichte herangezogen, die diese oben angeführten Kennwerte enthalten haben.

Folgende Semester wurden in die Berechnung mit einbezogen: WS 2012/2013, SS 2013, WS 2013/2014 und das SS 2014. Im SS 2013 und im SS 2014 haben Kennwerte aus dem dritten Semester gefehlt, im WS 2013/2014 haben Werte aus dem sechsten und achten Semester gefehlt. Diese Kennwerte wurden entweder unverändert oder nach Berechnung von Mittelwerten (über Semester hinweg) verwendet und dargestellt. In einer adaptierten Darstellung, die als solche gekennzeichnet ist, wurden nur Mittelwerte aus Semestern und Jahrgängen verwendet, bei denen mehr als 12 Studierende teilgenommen haben, um eine Verzerrung aufgrund einer zu geringen Teilnehmerzahl zu vermeiden.

5.2. STATISTISCHE METHODEN

Alle Bearbeitungen und Umformungen, die im Datenteil erwähnt wurden, wurden in Microsoft Excel (Microsoft Corp. 2010. Microsoft Office 2010. Redmond, Washington: Microsoft Corp.) durchgeführt. Dafür waren ausschließlich Standardfunktionen notwendig und bei sich wiederholenden Schritten wurden Makros verwendet. In gleicher Weise wurden für alle Kontrollschritte lediglich Standardfunktionen benutzt.

Die transformierten und bereinigten Tabellen wurden abschließend von allen Formatierungen bereinigt und in den Statistik-Paketen STATA (Stata Corp. 2015. Stata Statistical Software: Release 14. College Station, TX: Stata Corp LP.) oder SPSS (IBM 2015. SPSS Statistics: Release 22. Armonk, NY: IBM Corp.) weiterverarbeitet.

Für den Überblick über alle untersuchten Kategorien und Einheiten wurden absolute und prozentuelle Häufigkeiten verwendet und Mittelwerte berechnet. Auch bei den deskriptiven Beschreibungen der binären Fragetypen und Item-Eigenschaften und ordinalen Variablen der Distraktorenqualität wurden diese Kennwerte angegeben.

Bei den metrischen Variablen Item-Schwierigkeit und Item-Trennschärfe wurden neben Minima, Maxima, Mittelwerte und Standardabweichung auch Median-Werte angegeben. Um auf Normalverteilung zu prüfen, wurde der Kolmogorow-Smirnow-Test verwendet.

Bei der Berechnung der Gruppenunterschiede in Bezug auf Item-Schwierigkeit und Item-Trennschärfe wurden einerseits Mittelwerte, Mittelwertdifferenzen, Signifikanzwerte und die Effektstärken (Cohens d) angegeben, andererseits – bei den nichtparametrischen Verfahren – Mediane und die Signifikanzwerte. Zur Berechnung der Signifikanzwerte wurden der t-Test und der Mann-Whitney-U-Test herangezogen. Bei den Berechnungen des Einflusses der beiden ordinalen Variablen NFD-5% und NFD-0% wurde der Kruskal-Wallis-Test verwendet. Das Signifikanzniveau war bei allen Tests α=0,05 %.

Im Ergebnisteil werden vorwiegend Torten-, Balken- und Liniendiagramme verwendet. Es werden dabei einheitlich vier verschiedene Blautöne verwendet, lediglich bei Darstellung der PTM-Ergebnisse wird die Medizinische Universität Graz grün und die anderen Fakultäten/Universitäten blau dargestellt.

Vor dem Ergebnisteil sollen noch zwei Aspekte gesondert betont werden:

1.) Die binären Variablen teilen sich in Fragetypen (Typ-A positiv/negativ und Typ-K) und Item-Eigenschaften (Aussage, Vignette, Abbildung und Rechnung) auf. Die Fragetypen schließen einander aus und summieren sich auf 100 %. Die Item-Eigenschaften jedoch sind unabhängig voneinander, können auch gemeinsam vorkommen oder vollständig fehlen und ergeben daher keine Summe von 100 %.
2.) Im Gegensatz dazu werden im Ergebnisteil die metrischen Variablen Item-Schwierigkeit und Item-Trennschärfe als Kennwert bezeichnet.

6. Ergebnisse

6.1. Überblick über alle untersuchten Einheiten

Wie im Abschnitt Material abschließend beschrieben, wurden in Summe 90 Prüfungsdurchgänge und 4530 Fragenitems untersucht. Die nachfolgenden beiden Tabellen zeigen die Anzahl der Module, die Anzahl der Prüfungsdurchgänge, sowie die untersuchten Items und die Antritte pro Studienjahr und teilweise Abschnitt.

Jahr 2011/2012	Vorklinik	Zwischenklinik	Klinik	Gesamt
Anzahl der Module	5	4	4	13
Prüfungsdurchgänge (DG)	11	14	21	46
Untersuchte Items	558	484	1224	2266
Alle Items (DG x Itemzahl)				2352
Anteil der untersuchten Items				96,3 %
Antritte gesamt				4577
Mittelwert d. Antritte pro DG				99,5

Tabelle 2: Übersicht über alle untersuchten Module, Prüfungsdurchgänge, Items und Antritte im Studienjahr 2011/2012.

In Tabelle 2 ist zu sehen, dass 2011/2012 in den drei Abschnitten jeweils fünf, vier und vier Module, sowie 11, 14 und 21 Prüfungsdurchgänge untersucht wurden. In Summe wurden in diesem Jahr also 13 Module und 46 Prüfungsdurchgänge untersucht. Die Anzahl der untersuchten Items steigt von 558 und 484 Items in den ersten beiden Abschnitten auf 1224 im dritten Abschnitt an.

Der Grund liegt in den zahlreichen kleinen Prüfungsdurchgängen in diesem Abschnitt. Im Vergleich dazu gibt es in den ersten beiden Abschnitten wenige große Durchgänge. In Summe wurden 2266 Items untersucht, was 96,3 % aller möglichen Items entspricht. Die restlichen 3,7 % wurden aus diversen Gründen – wie im Abschnitt Material beschrieben – entfernt.

In diesem Jahr haben zudem bei allen untersuchten Modulen und Prüfungsdurchgängen 4577 Studierende teilgenommen, was einem Mittelwert von 99,5 Studierenden pro Prüfungsdurchgang entspricht (von 25 bis 334 Studierende).

2013/2014	Vorklinik	Zwischenklinik	Klinik	Gesamt
Anzahl der Module	5	4	4	13
Prüfungsdurchgänge (DG)	13	11	20	44
Untersuchte Items	687	377	1200	2264
Alle Items (DG x Itemzahl)				2268
Anteil der untersuchten Items				99,8 %
Antritte gesamt				4761
Mittelwert d. Antritte pro DG				108,2

Tabelle 3: Übersicht über alle untersuchten Module, Prüfungsdurchgänge, Items und Antritte im Studienjahr 2013/2014.

Parallel dazu ist in Tabelle 3 zu sehen, dass 2013/2014 in den drei Abschnitten ebenfalls fünf, vier und vier Module und 13, 11 und 20 Prüfungsdurchgänge untersucht wurden. In Summe wurden 13 Module und 44 Prüfungsdurchgänge untersucht. Die Anzahl der untersuchten Items pro Abschnitt ist im ersten Abschnitt 687, im Zweiten 377 und im Dritten 1200 (der Grund für die deutlich höhere Zahl im dritten Abschnitt ist derselbe, wie im Jahr 2011/2012). In Summe wurden mit 2264 Items beinahe gleich viele Items wie im Jahr 2011/2012 untersucht. Aufgrund der besseren Datenqualität wurden in diesem Jahr 99,8 % aller theoretisch möglichen Items untersucht. Im Jahr 2013/2014 haben bei allen untersuchten Modulen und Prüfungsdurchgängen 4761 Studierende teilgenommen, was einem Mittelwert von 108,2 Studierenden pro Prüfungsdurchgang entspricht (von 28 bis 344 Studierende).

6.2. VERTEILUNG DER ITEM-EIGENSCHAFTEN

In den nachfolgenden Tabellen werden alle Fragetypen und Item-Eigenschaften aufgelistet, die zwei Ausprägungen besitzen, also binär/dichotom verteilt sind. Alle Eigenschaften treffen entweder zu oder nicht zu.

Die Tabelle 4 zeigt die absoluten Häufigkeiten (Anzahl) und die relativen Häufigkeiten (Prozentangaben) der Items, aufgeteilt nach Studienjahren und Abschnitten. Die darauf folgende Tabelle 5 zeigt dieselben Kennwerte, allerdings für beide Studienjahre zusammen.

2011/2012	**Vorklinik**		**Zwischenklinik**		**Klinik**		**Gesamt**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Typ-A negativ	54	9,7	38	7,9	283	23,1	375	16,6
Typ-K	313	56,1	28	5,8	57	4,7	398	17,6
Aussage	215	38,5	49	10,1	243	19,9	507	22,4
Vignette	3	0,5	17	3,5	156	12,8	176	7,8
Abbildung	104	18,6	0	0,0	0	0,0	104	4,6
Rechnung	(in diesem Studienjahr nicht untersucht)							
2013/2014	**Vorklinik**		**Zwischenklinik**		**Klinik**		**Gesamt**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Typ-A negativ	87	12,7	25	6,6	368	30,7	480	21,2
Typ-K	338	49,2	108	28,7	55	4,6	501	22,1
Aussage	166	24,2	25	6,6	173	14,4	364	16,1
Vignette	3	0,4	3	0,8	80	6,7	86	3,8
Abbildung	124	18,1	5	1,3	1	0,1	130	5,7
Rechnung	34	5,0	6	1,6	1	0,1	41	1,8

Tabelle 4: Auflistung der Fragetypen und Item-Eigenschaften aller Abschnitte und beider Studienjahren, in absoluten und relativen Häufigkeiten.

Die Unterschiede zwischen dem ersten (2011/2012) und dem zweiten untersuchten Studienjahr (2013/2014) sind nicht besonders groß, zeigen jedoch die Variabilität der Eigenschaften. Zwischen dem Jahr 2011/2012 und dem Jahr 2013/2014 gibt es einen Anstieg an negativen Fragen (um 4,6 %) und eine Abnahme an Aussagen-basierten Fragen (um 6,3 %) und Vignettenfragen (um 4,0 %). In der Zwischenklinik gibt es einen etwas auffälligeren Anstieg an Fragen des Typs-K von 5,8 auf 28,7 %.

Einzig die Eigenschaft *Rechnung*, also ob eine Frage eine Rechnung im mathematischen Sinne enthält, wurde nur im zweiten Studienjahr untersucht.

Beide Jahre	**Vorklinik**		**Zwischenklinik**		**Klinik**		**Gesamt**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Typ-A negativ	141	11,3	63	7,3	651	26,9	855	18,9
Typ-K	651	52,3	136	15,8	112	4,6	899	19,9
Aussage	381	30,6	74	8,6	416	17,2	871	19,2
Vignette	6	0,5	20	2,3	236	9,7	262	5,8
Abbildung	228	18,3	5	0,6	1	0	234	5,2
Rechnung	34	5,0	6	1,6	1	0,1	41	1,8

Tabelle 5: Auflistung der Fragetypen und Item-Eigenschaften aller Abschnitte für beide Jahre gemeinsam, in absoluten und relativen Häufigkeiten.

Die folgenden Ausführungen beziehen sich auf beide Studienjahre: Negative Fragenformulierungen oder Fragen des Typs-A negativ kommen in allen Abschnitten, aber gehäuft im dritten Abschnitt vor (26,9 %). Fragen des Typs-K kommen vor allem in der Vorklinik vor (52,3 %) und nehmen im Laufe der drei Abschnitte deutlich ab. Aussagen-basierte Fragen kommen ebenfalls in allen Abschnitten zum Einsatz, am seltensten in der Zwischenklinik (8,6 %). Fragen mit Vignette kommen in den ersten beiden Abschnitten sehr selten vor (nur 6 und 20 Items), in der Klinik in 9,7 % der Fragen (236 Items). Umgekehrt kommen Abbildungen im ersten Abschnitt am häufigsten vor (18,3 %); bei Fragen mit Rechnungen ist das ähnlich (5,0 %), wenn auch in absoluten Zahlen seltener (34 Items mit Rechnungen und 228 Items mit Abbildungen).

In den oberen Tabellen wurden alle sechs Eigenschaften aufgelistet, in den beiden folgenden Balkendiagrammen werden jeweils nur drei Eigenschaften dargestellt. In Abbildung 15 ist der große Anteil an Fragen des Typs-K in der Vorklinik mit über 50 % gut zu sehen, negative Formulierungen kommen in der Klinik bei fast jeder dritten Frage zum Einsatz. Aber auch Aussagen-basierte Fragen haben in der Vorklinik einen Anteil von ca. 30 %.

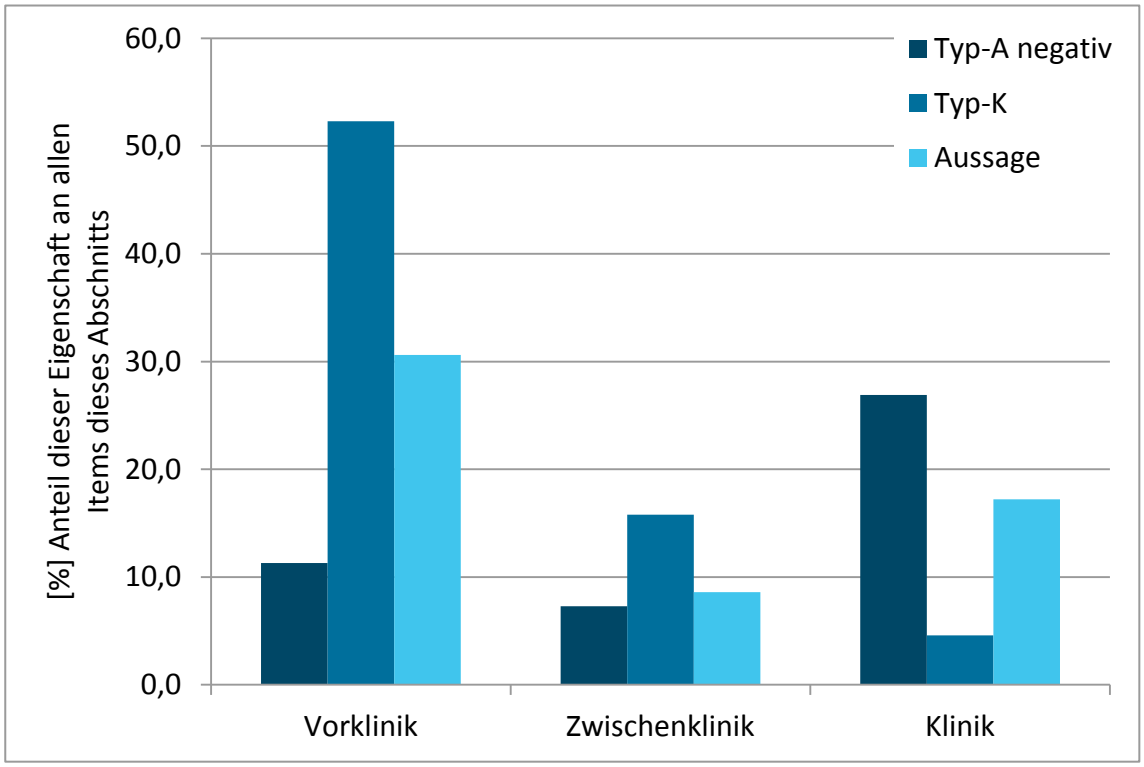

Abbildung 15: Grafische Darstellung des relativen Vorkommens der oben aufgelisteten Fragetypen und Item-Eigenschaften in den drei Abschnitten.

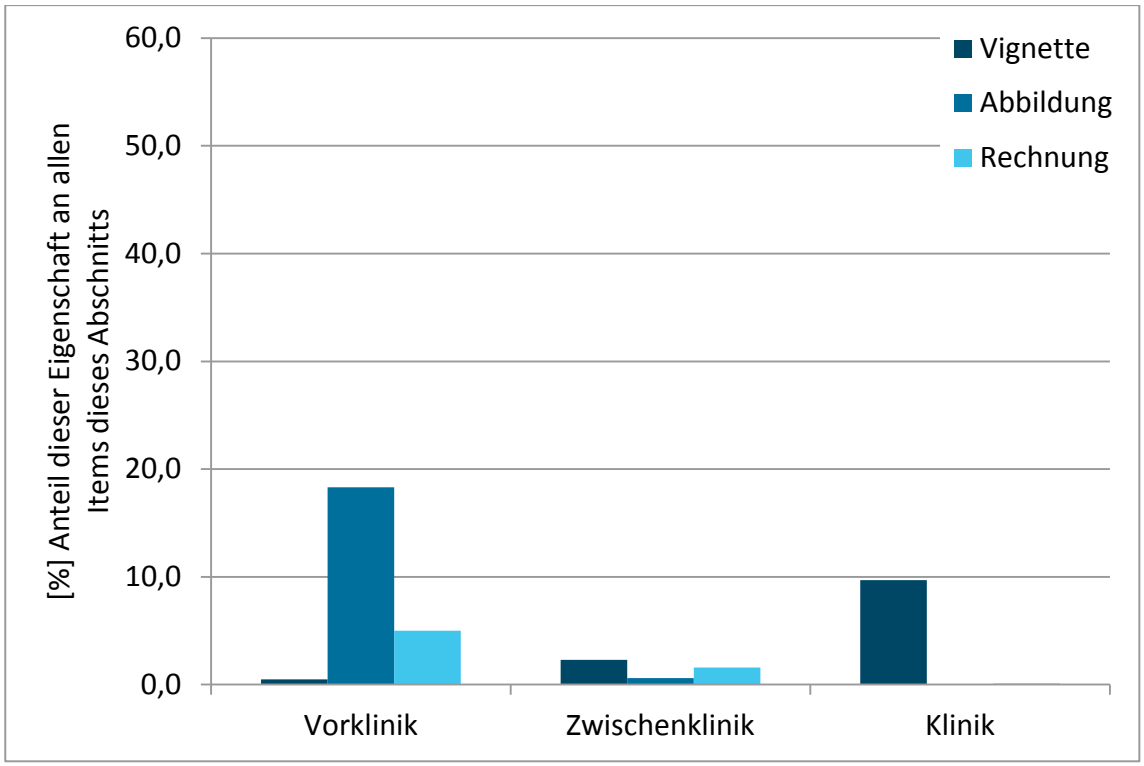

Abbildung 16: Grafische Darstellung des relativen Vorkommens der oben aufgelisteten Item-Eigenschaften in den drei Abschnitten.

Im Vergleich zu Abbildung 15 sind die Anteile in Abbildung 16 generell geringer, am meisten dominieren Fragen mit Abbildungen im vorklinischen Bereich (ca. 20 %). Fragen mit Fallvignetten kommen in Summe selten, am häufigsten jedoch im Bereich der Klinik vor (ca. 10 %).

In der nachfolgenden Abbildung 17 sind die vier Fragetypen Typ-A positiv/negativ und Typ-K positiv/negativ isoliert dargestellt. Der Typ-K hat einen Anteil am gesamten Pool von 19,8 %, der Anteil an negativen Fragen ist 18,8 %. Alle vier Fragetypen zusammen ergeben 100 %. Alle anderen Item-Eigenschaften wie Aussagen-basiert, Vignetten, Abbildungen oder Rechnungen sind separat zu sehen und kommen in allen möglichen Kombinationen vor. Diese Eigenschaften können sich daher nicht auf 100 % ergänzen.

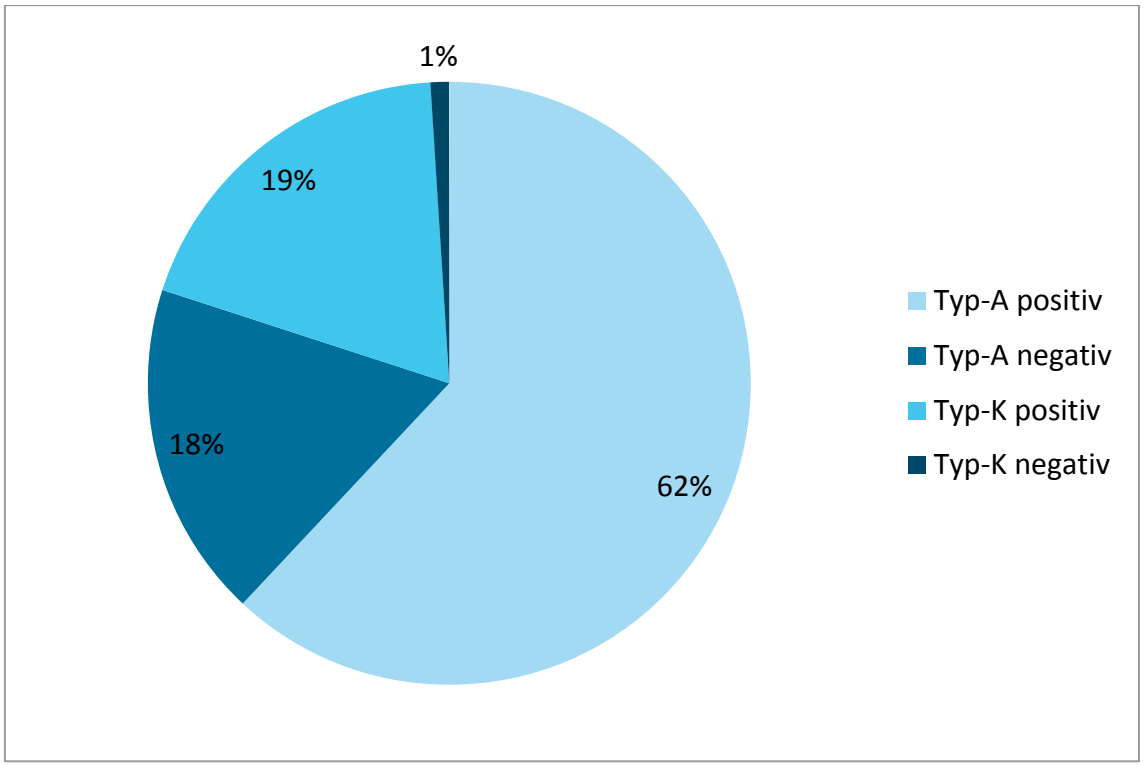

Abbildung 17: Tortendiagramm der Fragetypen, Typ-A und Typ-K, sowohl in positiven, wie in negativen Formulierungen.

Abschließend eine kleine Tabelle und ein kurzer Abschnitt zur Beschreibung, in welchen Kombinationen ausgewählte Fragetypen und Item-Eigenschaften vermehrt (Typ-A negativ, Typ-K, Aussage, Vignette und Abbildung) vorkommen:

	Abbildung	**Vignette**	**Aussage**	**Typ-K**
Typ-A neg.	- (seltener)	- (seltener)	+ (vermehrt)	- (seltener)
Typ-K	+ (vermehrt)	-- (nie)	+ (vermehrt)	
Aussage	+ (vermehrt)	-- (nie)		
Vignette	-- (nie)			

Tabelle 6: Raster zur Darstellung, in welchen Kombinationen ausgewählte Fragetypen und Item-Eigenschaften vorkommen.

Aus der Verteilung der Typen und Eigenschaften auf die einzelnen Abschnitte lässt sich bereits abschätzen, welche Kombinationen häufig sein könnten. Dies bestätigt sich auch hier: Typ-K, Aussagen und Abbildungen kommen vermehrt zusammen und im ersten Abschnitt vor.

Kombinationen aus den Eigenschaften Typ-K, Aussagen und Abbildungen zusammen mit Vignetten kommen nie vor. Ebenfalls selten sind Kombinationen aus Negativfragen und Fragen mit Abbildungen, Negativfragen und Vignetten, und Negativfragen und Fragen des Typs-K. Die einzige Ausnahme bildet hier die Kombination Aussagen-basierte Fragen und negative Formulierungen, diese Kombination ist überraschenderweise ebenfalls häufig.

6.3. Beschreibung der Variablen Schwierigkeit und Trennschärfe

In Tabelle 7 wird die metrische Variable Item-Schwierigkeit angeführt – zuerst getrennt für beide untersuchten Studienjahre und getrennt nach Abschnitten. Aufgrund der unterschiedlichen Verteilung wurden die Variablen Mittelwert und Standardabweichung mit dem Median ergänzt. Tabelle 8 zeigt dieselben Kennwerte für die Item-Trennschärfe.

2011/2012	**Vorklinik**	**Zwischenklinik**	**Klinik**	**Gesamt**
Minima	0,05	0,05	0,02	0,21
Maxima	1,00	1,00	1,00	1,00
Mittelwert	0,774	0,709	0,805	0,777
Standardabweichung	0,182	0,200	0,212	0,206
Median	0,825	0,741	0,881	0,836
2013/2014	**Vorklinik**	**Zwischenklinik**	**Klinik**	**Gesamt**
Minima	0,00	0,10	0,10	0,00
Maxima	1,00	1,00	1,00	1,00
Mittelwert	0,752	0,709	0,786	0,763
Standardabweichung	0,194	0,222	0,226	0,218
Median	0,797	0,752	0,752	0,828
Beide Jahre	**Vorklinik**	**Zwischenklinik**	**Klinik**	**Gesamt**
Minima	0,00	0,05	0,00	0,00
Maxima	1,00	1,00	1,00	1,00
Mittelwert	0,762	0,709	0,796	0,770
Standardabweichung	0,189	0,210	0,219	0,212
Median	0,811	0,750	0,875	0,833

Tabelle 7: Auflistung verschiedener Lage- und Streuungsmaße der Item-Schwierigkeiten, nach Studienjahren und Abschnitten.

In beiden Jahren, separat ebenso wie gemeinsam, gleichen sich die Minima- und Maxima-Angaben. Lediglich die Minima sind weder im Jahr 2011/2012 noch im zweiten und dritten Abschnitt im Jahr 2013/2014 null. Viele Fragen erreichen eine Item-Schwierigkeit von eins, jedoch nur bedingt eine Schwierigkeit von null. Bei ähnlicher Standardabweichung, nur die Vorklinik hat eine geringfügig kleinere, ergeben sich Schwierigkeitskennwerte von 0,762/0,709 und 0,796 über die drei Abschnitte hinweg. In den beiden getrennt dargestellten Studienjahren sieht die Situation nicht anders aus, der klinische Bereich hat die höchsten Schwierigkeitswerte (0,805 im ersten Jahr und 0,786 im zweiten), es handelt sich also um den Abschnitt mit den durchschnittlich leichtesten Fragen. Die zweitleichtesten Fragen hat die Vorklinik (0,774 im ersten Jahr und 0,752 im zweiten), die schwierigsten Fragen der zwischenklinische Bereich (0,709 in beiden Jahren).

2011/2012	**Vorklinik**	**Zwischenklinik**	**Klinik**	**Gesamt**
Minima	-0,14	-0,18	-0,33	-0,33
Maxima	0,80	0,77	0,71	0,80
Mittelwert	0,327	0,307	0,226	0,268
Standardabweichung	0,182	0,172	0,174	0,181
Median	0,316	0,316	0,238	0,275
2013/2014	**Vorklinik**	**Zwischenklinik**	**Klinik**	**Gesamt**
Minima	-0,35	-0,43	-0,43	-0,43
Maxima	0,82	0,79	0,79	0,82
Mittelwert	0,272	0,211	0,174	0,210
Standardabweichung	0,187	0,185	0,180	0,188
Median	0,269	0,202	0,202	0,207
Beide Jahre	**Vorklinik**	**Zwischenklinik**	**Klinik**	**Gesamt**
Minima	-0,35	-0,43	-0,43	-0,43
Maxima	0,82	0,79	0,74	0,82
Mittelwert	0,297	0,265	0,200	0,239
Standardabweichung	0,186	0,184	0,179	0,187
Median	0,290	0,273	0,204	0,243

Tabelle 8: Auflistung verschiedener Lage- und Streuungsmaße der Item-Trennschärfe, nach Studienjahren und Abschnitten.

Wenn man die Tabelle 8 bezogen auf die Item-Trennschärfe hernimmt, sieht man eine Schwankungsbreite von stark negativen Werten um -0,4 (alle Minima über Jahre und Abschnitte hinweg sind negativ), bis zu hoch positiven Werten deutlich über 0,5 (alle Maxima über Jahre und Abschnitte hinweg sind über 0,7). Unabhängig vom Studienjahr verschlechtert sich die durchschnittliche Trennschärfe im Laufe des Studiums, von 0,297 über 0,265 bis 0,200 (Mittelwerte beider Jahre).

In der folgenden Abbildung 18 sind diese Kennzahlen grafisch dargestellt. Man sieht die wechselnde Item-Schwierigkeit, mit dem höchsten Wert in der Klinik und die über die Abschnitte hinweg sinkende Item-Trennschärfe.

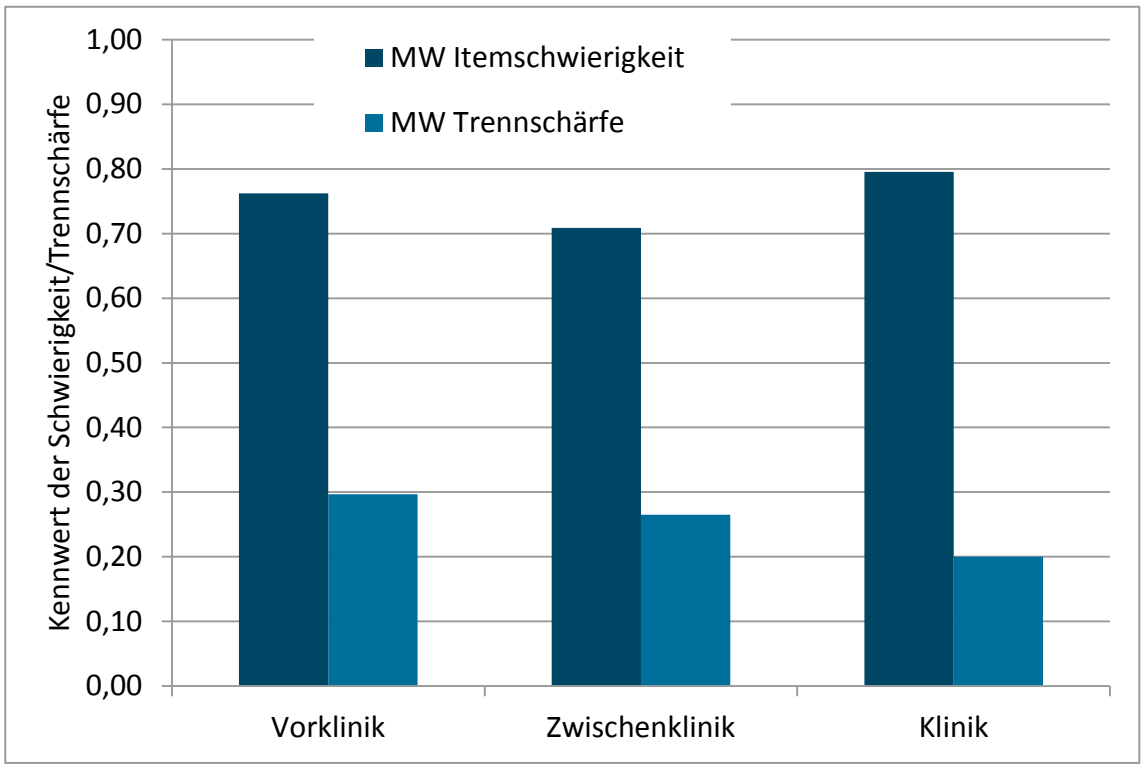

Abbildung 18: Grafische Darstellung der Entwicklung des Lagemaßes Mittelwert von Item-Schwierigkeit und Item-Trennschärfe über alle drei Abschnitte.

Die letzte Abbildung in diesem Kapitel (Abbildung 19) zeigt zwei Histogramme, die die Verteilung der beiden Kennwerte Item-Schwierigkeit und Item-Trennschärfe veranschaulicht. Das linke Histogramm zeigt die Verteilung der Item-Schwierigkeit, welche auffällig linksschief und nicht normalverteilt ist. Die Item-Schwierigkeit ist über alle drei Abschnitte hinweg nicht normalverteilt (p<0,001). Die schwarze vertikale Linie zeigt den Median des Kennwertes Item-Schwierigkeit, welche mit 0,833 deutlich über 0,8 liegt und das Dominieren sehr leichter Fragenitems bestätigt. Das rechte Histogramm zeigt die Verteilung der Item-Trennschärfe, welche optisch annähernd normalverteilt aussieht. In der Vor- und Zwischenklinik liegt eine Normalverteilung vor (p=0,200 und p=0,188), jedoch nicht im klinischen Abschnitt (p<0,001). Auch hier wurde eine vertikale Linie beim Median gesetzt, dieser liegt bei 0,290. Auffällig ist bei der rechten Abbildung die extreme Häufigkeit bei der Trennschärfe null. Alle Fragenitems, bei der die Schwierigkeit entweder null oder eins ist, besitzen rechentechnisch und real eine Trennschärfe von null.

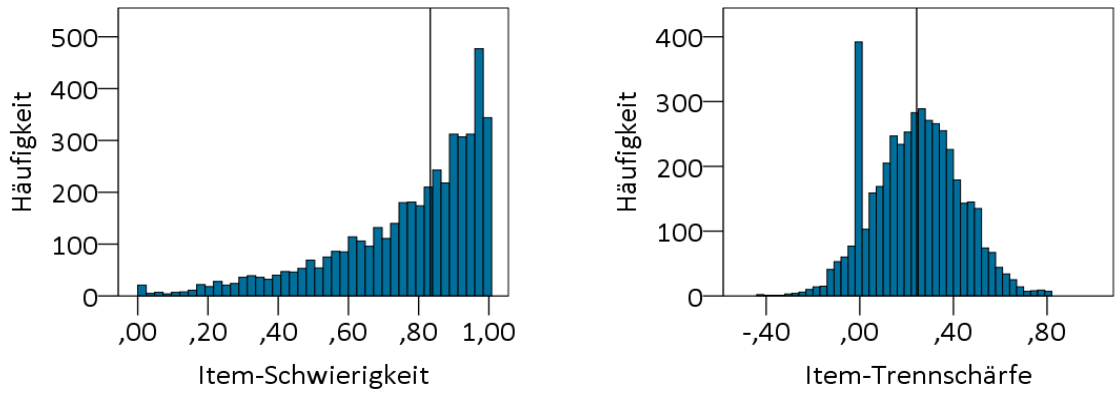

Abbildung 19: Zwei Histogramme, die die Item-Schwierigkeit (links) und die Item-Trennschärfe (rechts) zeigen. Die beiden vertikalen Linien markieren die Median-Werte.

6.4. DIE ITEM-SCHWIERIGKEIT IN KATEGORIEN

Im Kapitel 6.3 werden die beiden metrischen Variablen Item-Schwierigkeit und Item-Trennschärfe über Mittelwerte und Mediane beschrieben. Um ein genaueres Bild zu erhalten und Vergleiche mit der Literatur zu ermöglichen, wird die Item-Schwierigkeit nochmals über die Kategorien <0,2 sowie 0,2 bis 0,8 und >0,8 beschrieben. Dies wird in Abbildung 20 für die Abschnitte und in Abbildung 21 für die Fragetypen und Item-Eigenschaften gemacht.

Dabei zeigt sich für die Abschnitte, dass die Klinik die meisten leichten Items mit einer Schwierigkeit >0,8 hat (63 %), gefolgt von der Vorklinik (52 %) und Zwischenklinik (40 %). Entsprechend verändert sich der Anteil an mittelschweren Fragen (0,2 bis 0,8), der Anteil der schweren Fragen (>0,8) ist in allen drei Abschnitten sehr klein (1 %, 2 % und 3 %).

Wenn man sich die ersten vier Eigenschaften in Abbildung 21 ansieht – Abbildungen und Rechnungen aufgrund des geringen Vorkommens ausgenommen – dann hat die Eigenschaft *Vignette* die leichtesten Fragen: hoher Anteil an leichten Fragen mit Schwierigkeit >0,8 (72 %) und entsprechend kleiner Anteil an mittelschweren Fragen (27 %). Schwere Fragen gibt es in der Kategorie *Vignette* nur zwei. Der zweitleichteste Fragentyp ist die *negative Frage* mit 57 % leichten und 41 % mittelschweren Fragen. Die Eigenschaften *Typ-K* und *Aussage* gehen mit schwereren Fragen einher.

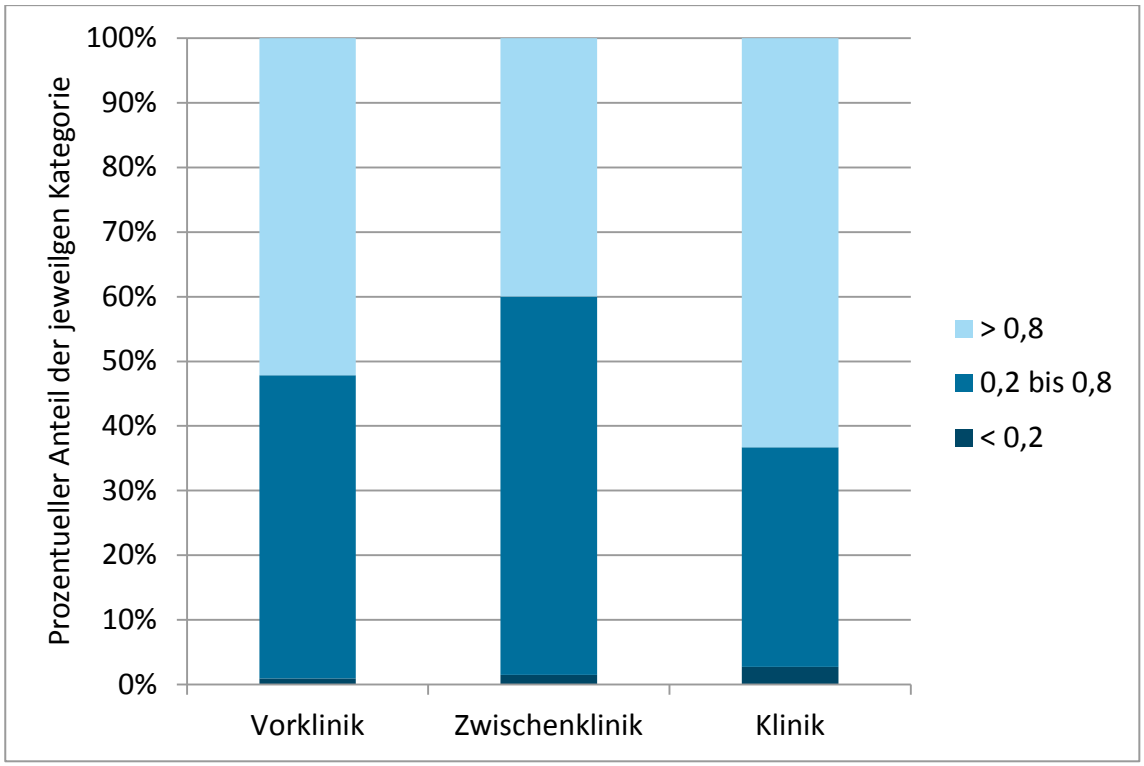

Abbildung 20: Der Anteil an leichten (>0,8), mittelschweren (0,2 bis 0,8) und schweren (<0,2) Fragen in Abhängigkeit der Abschnitte.

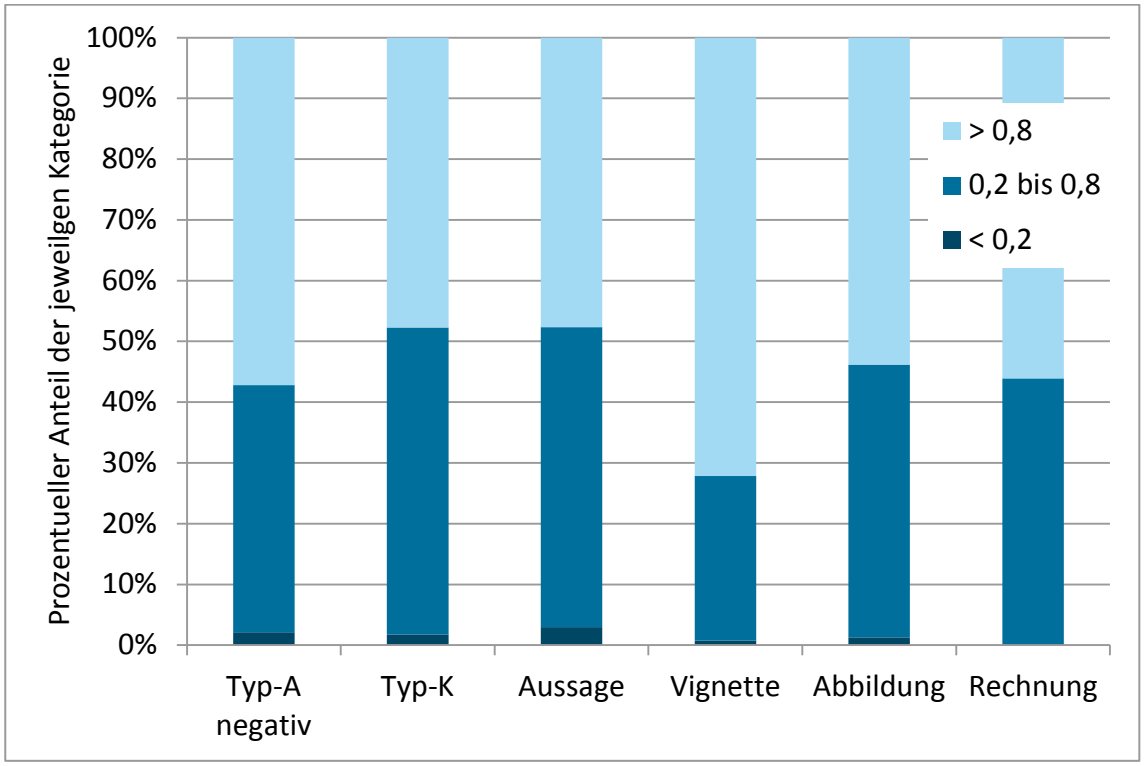

Abbildung 21: Der Anteil an leichten (>0,8), mittelschweren (0,2 bis 0,8) und schweren (<0,2) Fragen in Abhängigkeit der Fragetypen und Item-Eigenschaften.

6.5. VERTEILUNG DER EIGENSCHAFT DES DOMINIERENDEN DISTRAKTORS

Wiederholend sei erwähnt, dass *DiDo* (kurz für *Distraktor dominiert*) eine für diese Arbeit eingeführte künstliche Bezeichnung ist. Sie bedeutet, dass ein Distraktor, also eine inkorrekte Antwortoption häufiger gewählt wurde, als die korrekte Antwortoption.

Beide Jahre	Vorklinik		Zwischenklinik		Klinik		Gesamt	
	N	%	N	%	N	%	N	%
DiDo 2011/2012	23	4,1	27	5,6	66	5,4	116	5,1
DiDo 2013/2014	18	2,6	39	10,3	76	6,3	133	5,9
DiDo beide Jahre	41	3,3	66	7,7	142	5,9	249	5,5

Tabelle 9: Auflistung der Variable DiDo (*Distraktor dominiert*) in allen Abschnitten und in beiden Studienjahren, in absoluten und relativen Häufigkeiten.

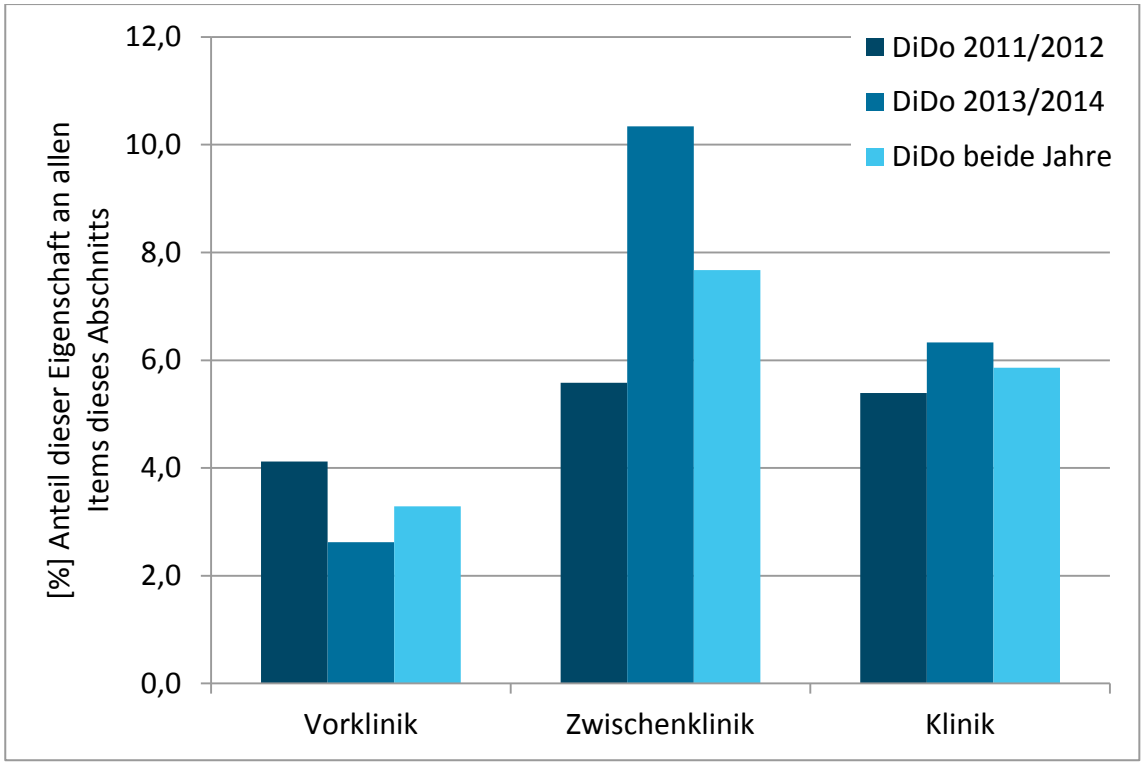

Abbildung 22: Grafische Darstellung des Vorkommens der dominierenden Distraktoren über alle drei Abschnitte in beiden Jahren bzw. zusammen.

Sowohl Tabelle 9, wie auch Abbildung 22 zeigen die Verteilung der Item-Eigenschaft *DiDo*. Das Vorkommen dieser unerwünschten Eigenschaft liegt zwischen 2,6 % und 10,3 % aller Fragen in einem Abschnitt, schwankt, wenn man die Studienjahre miteinander vergleicht, und ist im Bereich der Zwischenklinik mit 7,7 % am häufigsten. Am zweithäufigsten kommen diese Fragen in der Klinik vor (5,9 %), am seltensten in der Vorklinik (3,3 %).

6.6. Verteilung der Kennzahlen zur Qualität der Distraktoren

Es gibt in den Datensätzen zwei Variablen, die ordinal skaliert sind, nämlich *NFD-5%* und *NFD-0%*, erstere bedeutet die Anzahl (N) der Antwortoptionen (AWO), die von weniger als 5 % der Studierenden (u5P) dieses Prüfungsdurchganges gewählt wurde, die zweite Variable die Anzahl der Antwortoptionen, die von niemanden (null) gewählt wurde. Schnell erläutert, stehen diese Kennzahlen für *schlechte* oder *sehr schlechte* falsche Antwortoptionen bzw. Distraktoren.

NFD-5%	**Vorklinik**		**Zwischenklinik**		**Klinik**	
	N	**%**	**N**	**%**	**N**	**%**
keine <5%	101	8,1	109	12,7	161	6,6
eine <5%	189	15,2	172	20,0	248	10,2
zwei <5%	298	23,9	200	23,2	424	17,5
drei <5%	348	28,0	208	24,2	663	27,4
vier <5%	307	24,7	170	19,7	927	38,2
fünf <5%	2	0,2	2	0,2	1	0,0
NFD-0%	**Vorklinik**		**Zwischenklinik**		**Klinik**	
	N	**%**	**N**	**%**	**N**	**%**
keine =0	676	54,3	465	54,0	590	24,3
eine =0	293	23,5	213	24,7	613	25,3
zwei =0	176	14,1	116	13,5	537	22,2
drei =0	82	6,6	59	6,9	427	17,6
vier =0	18	1,5	8	0,9	257	10,6
fünf =0	0	0,0	0	0,0	0	0,0

Tabelle 10: NFD-5% und NFD-0%: Absolute und relative Häufigkeiten pro Abschnitt.

Die häufigste Kategorie bei der Kennzahl NFD-5% in der Vorklinik hat einen Anteil von 28 % und dabei sind drei Antwortoptionen sehr selten gewählt. Dasselbe Bild zeigt sich in der Zwischenklinik, bei der in 24,2 % der Fragen drei Optionen schlecht sind. In der Klinik sind es mit 38,2 % sogar vier schlechte Optionen. Da die meisten Fragen nur fünf Antwortoptionen und damit vier Distraktoren haben, bedeutet das, dass in der Klinik ein Drittel aller Fragen keine guten Distraktoren haben.

Wenn man sich die Kennzahl NFD-0% im zweiten Teil der Tabelle 10 ansieht, ist die Verteilung weniger deutlich. Die häufigste Ausprägung (54,3 und 54,0 %) ist in der Vor- und Zwischenklinik die, bei der keine Antwortoption von keinem der Studierenden gewählt worden ist. Hingegen kommen in der Klinik auch öfter mehrere sehr schlechte Antwortoptionen vor: In 25,3 % der Fragen ist ein Distraktor nicht gewählt, in 22,2 % sind es zwei und in immerhin noch 10,6 % sind es alle vier Distraktoren, die nicht gewählt worden sind.

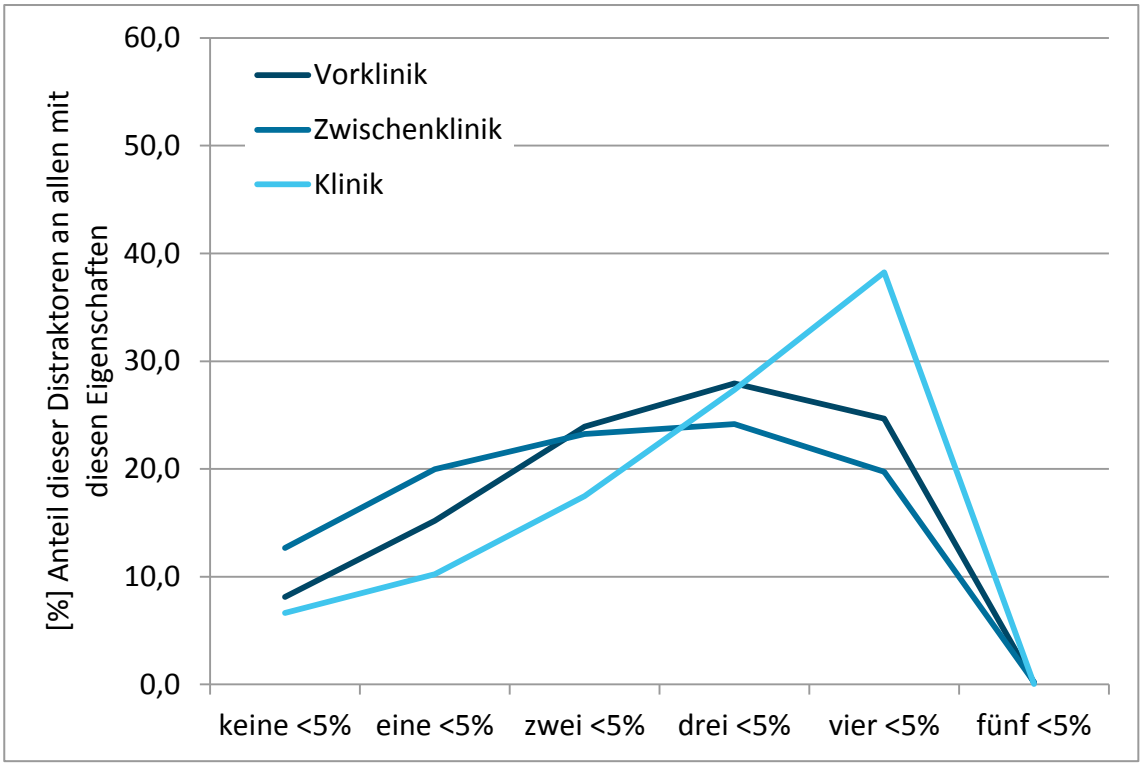

Abbildung 23: Grafische Darstellung der Verteilung „schlechter“ Distraktoren (von weniger als 5 % der Studierenden gewählt).

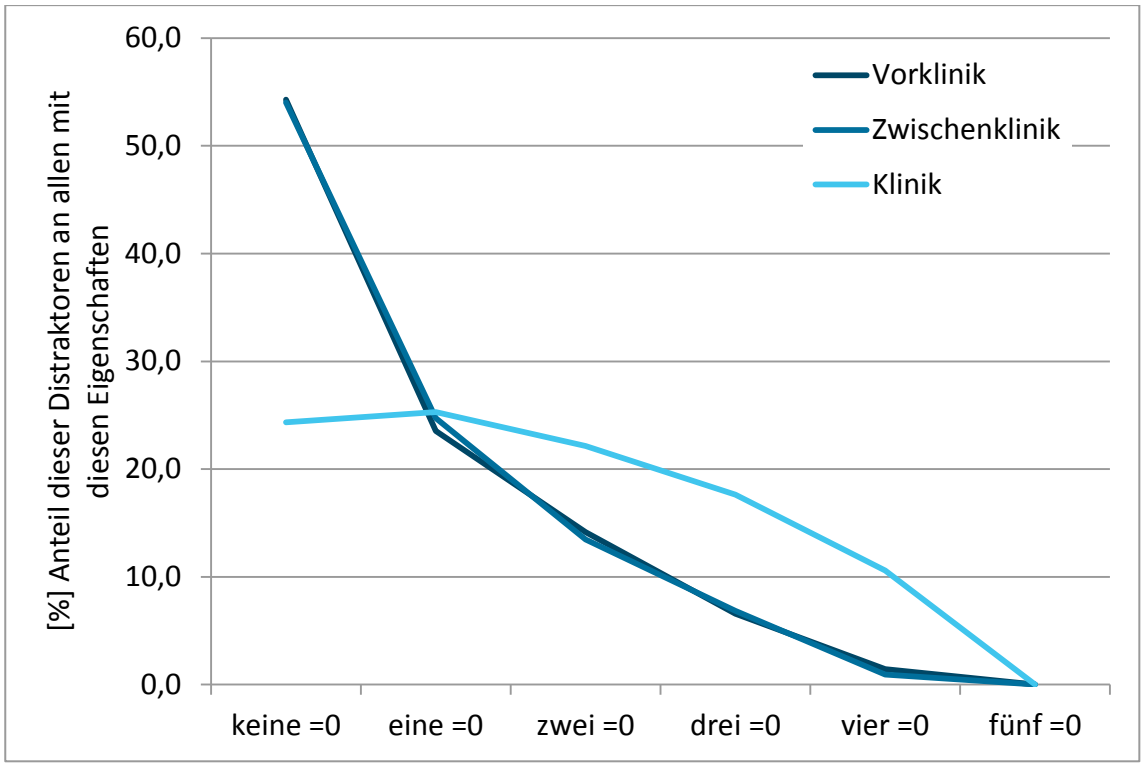

Abbildung 24: Grafische Darstellung der Verteilung „ganz schlechter“ Distraktoren (von keiner/keinem der Studierenden gewählt).

In Abbildung 23 und Abbildung 24 ist die Verteilung der Anzahl der schlechten (NFD-5%) und sehr schlechten (NFD-0%) Antwortoptionen grafisch dargestellt. Zwischen der Vor- und Zwischenklinik gibt es wenig Unterschied, allerdings sind die wenig oder nicht gewählten Antwortoptionen in der Klinik besonders häufig.

In Abbildung 25 werden die Kategorien null bis fünf schlechte Antworten pro Fragenitem vereinfacht: es wird der Mittelwert der Kennzahlen NFD-5% und NFD-0% berechnet und angezeigt. Der Mittelwert von NFD-5% ist mit 2,9 in der Klinik am größten, gefolgt von der Vorklinik mit 2,5 und der Zwischenklinik mit 2,3. Der Wert 2,9 bedeutet, dass knapp drei der meist vier Distraktoren schlecht sind und von weniger als 5 % der Studierenden gewählt wurde. Der Mittelwert von NFD-0% ist ebenfalls mit 1,7 mehr als doppelt so hoch wie in der Vor- oder Zwischenklinikmit je 0,8.

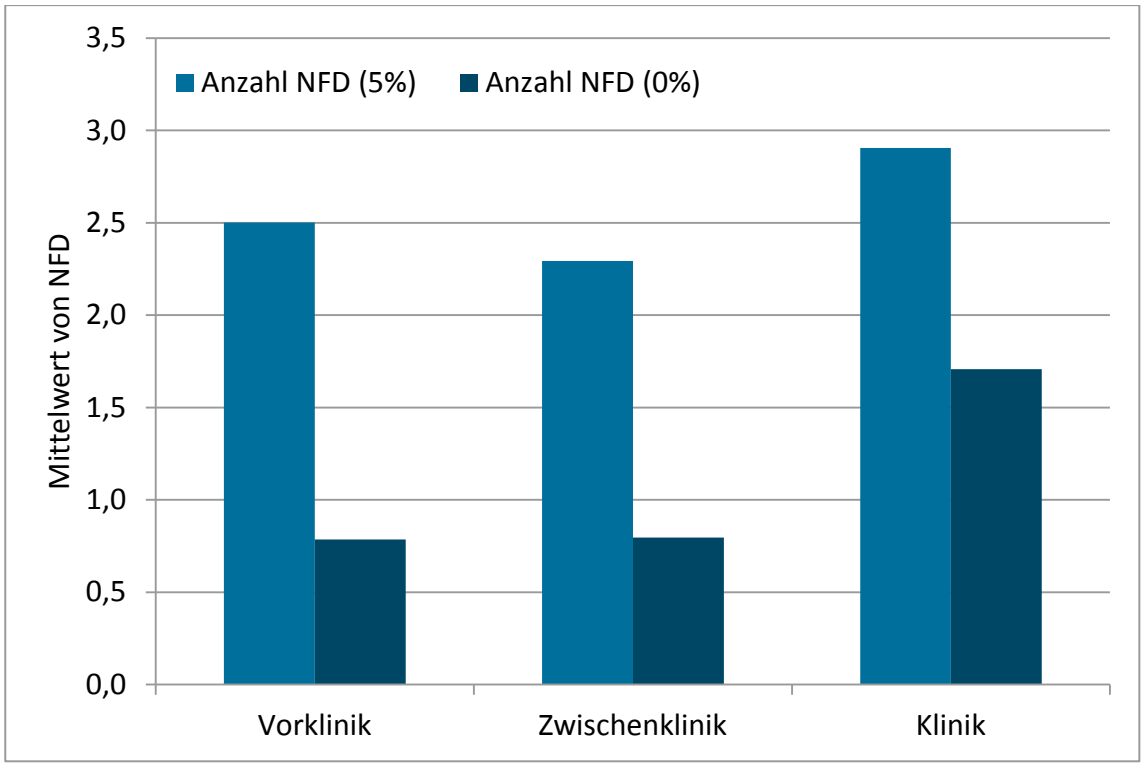

Abbildung 25: Balkendiagramm zur Darstellung von NFD-5% und NFD-0% in den drei Abschnitten.

6.7. Gruppenunterschiede in Bezug auf Schwierigkeit und Trennschärfe

In diesem Abschnitt werden die Lagemaße der einzelnen Eigenschaften, je nachdem, ob die Eigenschaft zutrifft (*pos*) oder nicht (*neg*), miteinander verglichen. In beiden Jahren und zusammen für beide Jahre werden die Mittelwerte beschrieben und auf einen signifikanten Unterschied hin getestet. Im Anschluss werden die Unterschiede in einem nichtparametrischen Verfahren berechnet und die Effektstärken Cohens d angegeben.

2011/2012	N Total	pos (MW)	neg (MW)	MW Diff	t	p (t-Test)		p (MWU)	d
Typ-A neg.	375	0,80	0,77	-0,031	-2,75	0,006	*	0,004	-0,15
Typ-K	398	0,76	0,78	0,018	1,77	0,078		<0,001	0,09
Aussage	507	0,72	0,79	0,078	7,09	<0,001	*	<0,001	0,38
Vignette	176	0,85	0,77	-0,076	-5,92	<0,001	*	<0,001	-0,07
Abbildung	104	0,80	0,78	-0,023	-1,50	0,136		0,677	-0,11
2013/2014	**N Total**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	480	0,76	0,76	0,004	0,36	0,716		0,918	0,02
Typ-K	501	0,73	0,77	0,046	4,32	<0,001	*	<0,001	0,21
Aussage	364	0,75	0,77	0,016	1,26	0,207		0,074	0,07
Vignette	86	0,83	0,76	-0,066	-3,23	0,002	*	0,004	-0,30
Abbildung	130	0,76	0,76	0,004	0,20	0,841		0,165	0,02
Rechnung	41	0,77	0,76	-0,006	-0,17	0,863		0,949	-0,03
Beide Jahre	**N Total**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	855	0,78	0,77	-0,011	-1,31	0,191		0,074	-0,05
Typ-K	899	0,74	0,78	0,034	4,59	<0,001	*	<0,001	0,16
Aussage	871	0,73	0,78	0,049	5,92	<0,001	*	<0,001	0,23
Vignette	262	0,84	0,77	-0,075	-6,87	<0,001	*	<0,001	-0,35
Abbildung	234	0,78	0,77	-0,007	-0,61	0,540		0,174	-0,04
Rechnung	41	0,77	0,76	-0,006	-0,17	0,863		0,949	-0,03

Tabelle 11: Deskriptive Kennzahlen, Kennzahlen der Tests und die Effektstärke d der Item-Schwierigkeit in Abhängigkeit der binären Eigenschaften.

Wenn man die Zahlen im dritten Teil der Tabelle 11 – für beide Studienjahre zusammen berechnet – ansieht, sieht man drei signifikante Unterschiede: Fragen des Typs-K sind deutlich schwerer (MW-Differenz 0,034 und p<0,001), Aussagen-basierte Fragen ebenso (MW-Differenz 0,049 und p<0,001), dagegen sind Fragen mit Vignetten deutlich leichter (MW-Differenz -0,075 und p<0,001).

Für beide Studienjahre getrennt berechnet, können nicht immer alle signifikanten Unterschiede bestätigt werden. Im ersten Jahr bestätigen sich die Differenzen betreffend Aussagen-basierten- und Vignetten-Fragen, im zweiten Jahr betreffend Typ-K und Vignette. Zusätzlich findet sich im ersten Studienjahr noch ein weiterer signifikanter Unterschied, denn negativ formulierte Fragen sind hier leichter.

Ebenfalls in diesem Jahr gibt es bei Fragen des Typs-K einen signifikanten Unterschied, welcher sich jedoch nur im nichtparametrischen Test (Mann-Whitney-U) bestätigt.

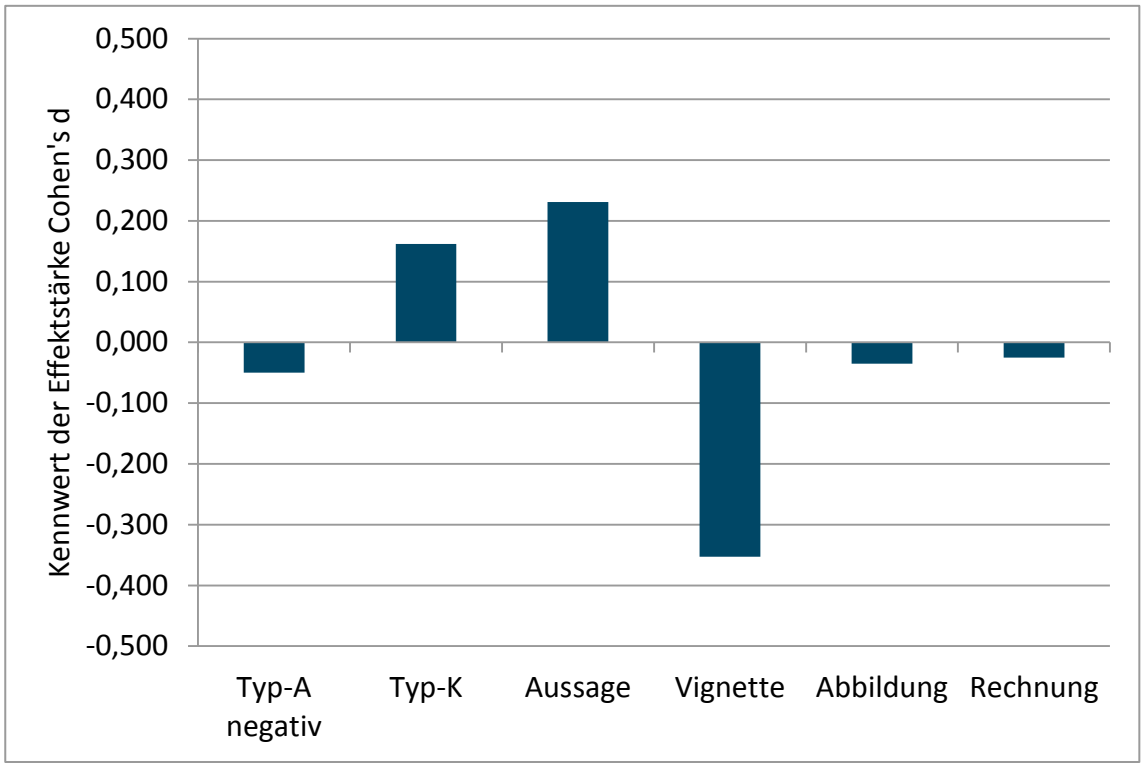

Abbildung 26: Grafische Darstellung der Effektstärken der Item-Schwierigkeiten in Abhängigkeit der unterschiedlichen Fragetypen und Item-Eigenschaften. Hinweis: Positive Zahlen und Balken nach oben kennzeichnen schwerere Fragenitems, negative Zahlen und Balken nach unten leichtere Fragenitems.

Grob zusammengefasst zeigt das Ergebnis, wenn man sich die grafische Darstellung der Effektstärken d in Abbildung 26 ansieht, deutlich schwerere Fragenitems kommen in Zusammenhang mit Typ-K Fragen (d=0,162) und in Zusammenhang mit Aussagen-basierten Fragen (d=0,231) vor, deutlich leichter sind die Fragen, wenn Fallvignetten vorkommen (d=-0,353).

2011/2012	N Total	pos (MW)	neg (MW)	MW Diff	t	p (t-Test)		p (MWU)	d
Typ-A neg.	375	0,26	0,27	0,011	1,16	0,245		0,374	0,06
Typ-K	398	0,32	0,26	-0,067	-6,30	<0,001	*	<0,001	-0,37
Aussage	507	0,30	0,26	-0,042	-4,50	<0,001	*	<0,001	-0,23
Vignette	176	0,24	0,27	0,033	2,27	0,024	*	0,028	0,18
Abbildung	104	0,36	0,26	-0,097	-4,99	<0,001	*	<0,001	-0,54
2013/2014	**N Total**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	480	0,20	0,21	0,013	1,31	0,191		0,160	0,07
Typ-K	501	0,24	0,20	-0,039	-3,96	<0,001	*	<0,001	-0,21
Aussage	364	0,21	0,21	-0,003	-0,29	0,771		0,629	-0,02
Vignette	86	0,19	0,21	0,022	1,14	0,255		0,270	0,12
Abbildung	130	0,30	0,21	-0,093	-5,43	<0,001	*	<0,001	-0,50
Rechnung	41	0,17	0,21	0,037	1,76	0,086		0,182	0,20
Beide Jahre	**N Total**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	855	0,23	0,24	0,017	2,35	0,019	*	0,033	0,09
Typ-K	899	0,28	0,23	-0,047	-6,44	<0,001	*	<0,001	-0,25
Aussage	871	0,26	0,23	-0,037	-4,34	<0,001	*	<0,001	-0,16
Vignette	262	0,22	0,24	0,019	1,59	0,114		0,119	0,10
Abbildung	234	0,33	0,23	-0,091	-7,03	<0,001	*	<0,001	-0,49
Rechnung	41	0,17	0,21	0,037	1,76	0,086		0,182	0,20

Tabelle 12: Deskriptive Kennzahlen, Kennzahlen der Tests und die Effektstärke d der Item-Trennschärfe in Abhängigkeit der binären Eigenschaften.

Betreffend die Item-Trennschärfe (wie aus der Tabelle 12 entnommen) wurden folgende signifikanten Unterschiede bestätigt: Negative Formulierungen sind weniger trennscharf (0,23 statt 0,24 und p=0,019), dagegen sind Fragen des Typs-K (0,28 statt 0,23 und p<0,001), Fragen mit Aussagen (0,26 statt 0,23 und p<0,001) oder mit Abbildungen (0,33 statt 0,23 und p<0,001) deutlich trennschärfer. Die Richtung der Unterschiede zeigt sich auch in den getrennten Berechnungen, allerdings sind diese Unterschiede nicht immer signifikant.

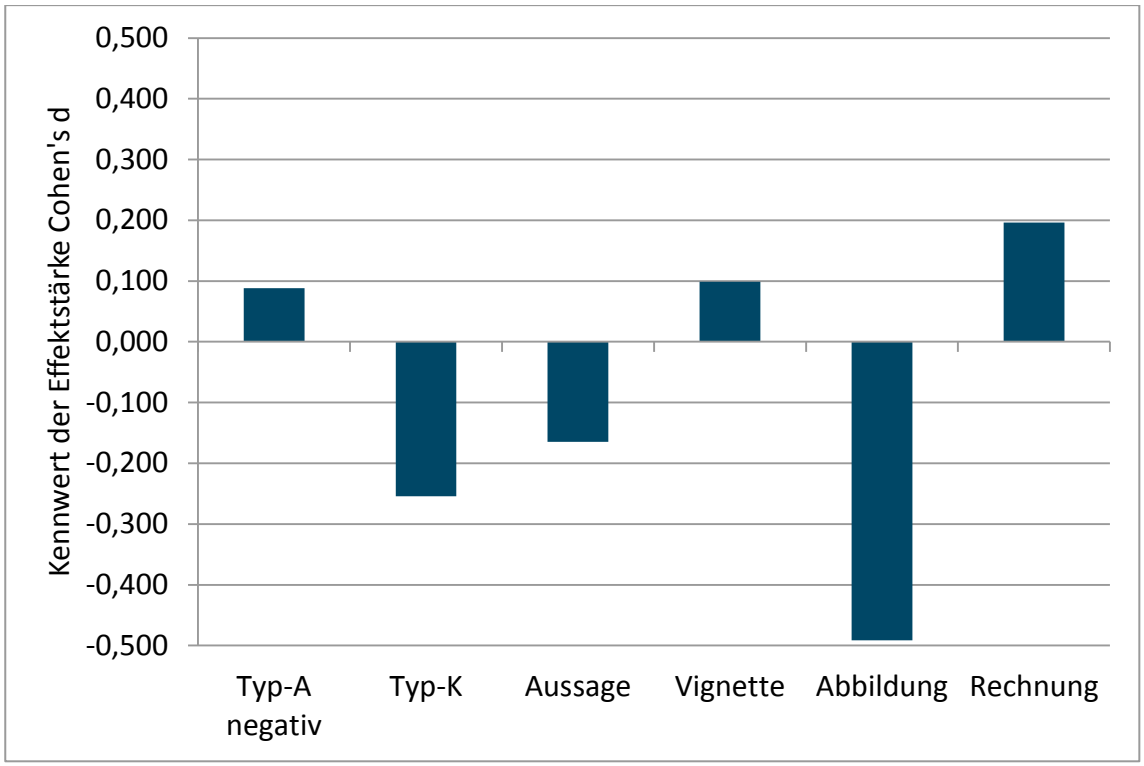

Abbildung 27: Grafische Darstellung der Effektstärken der Item-Trennschärfen in Abhängigkeit der unterschiedlichen Fragetypen und Eigenschaften. Hinweis: Negative Zahlen und Balken nach unten kennzeichnen mehr trennscharfe Fragenitems, positive Zahlen und Balken nach oben weniger trennscharfe Fragenitems.

Wiederum grob zusammengefasst wird das Ergebnis, wenn man sich die grafische Darstellung der Effektstärken d in Abbildung 27 ansieht. Typ-A negative Fragen sind weniger trennscharf, wenn auch mit geringer Effektstärke (d=0,09). Typ-K (d=-0,254) und Aussagen-basierte Fragen (d=-0,165), ebenso wie Fragen mit Abbildungen (d=-0,492) sind deutlich trennschärfer, auch mit signifikantem Unterschied. Vignetten-Fragen zeigen nicht in beiden Jahren einen signifikanten Unterschied (0,024 im ersten Jahr und 0,255 im zweiten Jahr), Fragen mit Rechnungen sind insgesamt zu selten, um belastbare Aussagen treffen zu können.

Vorklinik	N klGr	pos (MW)	neg (MW)	MW Diff	t	p (t-Test)		p (MWU)	d
Typ-A neg.	141	0,75	0,76	0,011	0,61	0,273		0,716	0,06
Typ-K	594	0,75	0,77	0,018	1,68	0,046	*	0,001	0,10
Aussage	381	0,76	0,76	0,003	0,28	0,389		0,387	0,02
Vignette	6	0,86	0,76	-0,096	-2,75	0,019	*	0,241	-0,51
Abbildung	228	0,78	0,76	-0,026	-2,04	0,021	*	0,191	-0,14
Rechnung	34	0,77	0,75	-0,021	-0,65	0,261		0,497	-0,11
Zwischen-klinik	**N klGr**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	63	0,69	0,71	0,020	0,78	0,218		0,272	0,10
Typ-K	136	0,72	0,71	-0,009	-0,43	0,334		0,188	-0,05
Aussage	74	0,68	0,71	0,032	1,17	0,122		0,262	0,15
Vignette	20	0,79	0,71	-0,087	-3,04	0,003	*	0,113	-0,42
Abbildung	5	0,51	0,71	0,195	1,59	0,093		0,092	0,93
Rechnung	6	0,71	0,71	0,003	0,02	0,491		0,825	0,01
Klinik	**N klGr**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	651	0,79	0,80	0,004	0,42	0,338		0,227	0,02
Typ-K	112	0,71	0,80	0,090	3,93	0,000	*	0,000	0,41
Aussage	416	0,71	0,81	0,100	7,70	0,000	*	0,000	0,46
Vignette	236	0,84	0,79	-0,053	-4,37	0,000	*	0,005	-0,24
Abbildung	1	0,59	0,80	0,203				0,247	
Rechnung	1	1,00	0,79	-0,214				0,115	

Tabelle 13: Deskriptive Kennzahlen, Kennzahlen der Tests und die Effektstärke d der Item-Schwierigkeiten in Abhängigkeit der binären Item-Eigenschaften – nach Abschnitten.

Wenn man sich den Einfluss auf die Item-Schwierigkeit getrennt nach Abschnitten ansieht in Tabelle 13, sieht man den schwierigeren Typ-K in der Vorklinik (d=0,09) und der Klinik (d=0,41), den schwierigeren Aussagen-basierten Typ ebenfalls in der Klinik (d=0,46), die schwierigeren Fragen mit Abbildungen nur in der Vorklinik (d=-0,14). Beim letzten Punkt muss man ergänzen, dass sich nur in der Vorklinik eine entsprechende Anzahl mit diesem Fragentyp befindet (228 Items). Fragen mit Vignetten sind über alle Abschnitte hinweg leichter (d -0,51/-0,42/-0,24), allerdings sind in den beiden ersten Abschnitten ebenfalls nur wenige Items untersucht worden (6 und 20 Items).

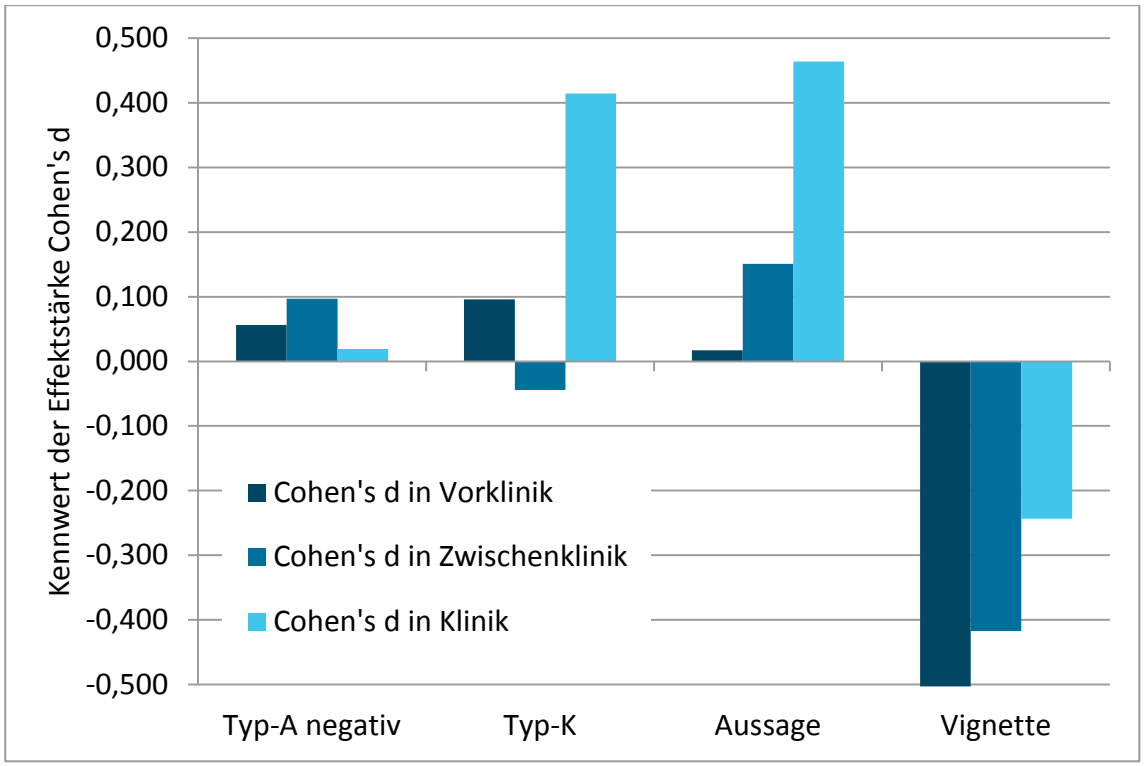

Abbildung 28: Grafische Darstellung der Effektstärken der Item-Schwierigkeiten in Abhängigkeit der unterschiedlichen Fragetypen und Item-Eigenschaften. Hinweis: Positive Zahlen und Balken nach oben kennzeichnen schwerere Fragenitems, negative Zahlen und Balken nach unten leichtere Fragenitems.

Diese Ergebnisse – Fragen des Typs-K und mit Aussagen sind schwerer und Vignettenfragen sind leichter – zeigen sich auch in den dargestellten Effektstärken in Abbildung 28.

Vorklinik	N klGr	pos (MW)	neg (MW)	MW Diff	t	p (t-Test)		p (MWU)	d
Typ-A neg.	141	0,34	0,29	-0,044	-2,69	0,004	*	0,001	-0,24
Typ-K	594	0,32	0,28	-0,041	-3,90	0,000	*	0,001	-0,22
Aussage	381	0,31	0,29	-0,024	-1,99	0,023	*	0,037	-0,13
Vignette	6	0,25	0,30	0,048	0,87	0,211		0,435	0,26
Abbildung	228	0,33	0,29	-0,042	-2,97	0,002	*	0,001	-0,22
Rechnung	34	0,20	0,28	0,081	3,50	0,001	*	0,006	0,44
Zwischen-klinik	**N klGr**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	63	0,28	0,26	-0,013	-0,63	0,267		0,653	-0,07
Typ-K	136	0,13	0,29	0,163	12,20	0,000	*	0,000	0,93
Aussage	74	0,25	0,27	0,013	0,64	0,261		0,347	0,07
Vignette	20	0,29	0,26	-0,023	-0,53	0,300		0,827	-0,13
Abbildung	5	0,16	0,27	0,103	3,31	0,012	*	0,129	0,56
Rechnung	6	0,08	0,21	0,131	2,93	0,014	*	0,053	0,71
Klinik	**N klGr**	**pos (MW)**	**neg (MW)**	**MW Diff**	**t**	**p (t-Test)**		**p (MWU)**	**d**
Typ-A neg.	651	0,20	0,20	0,005	0,56	0,287		0,499	0,03
Typ-K	112	0,23	0,20	-0,031	-1,75	0,041	*	0,043	-0,17
Aussage	416	0,22	0,20	-0,025	-2,64	0,004	*	0,004	-0,14
Vignette	236	0,22	0,20	-0,017	-1,35	0,089		0,223	-0,09
Abbildung	1	-0,04	0,20	0,239				0,133	
Rechnung	1	0,00	0,17	0,174				0,262	

Tabelle 14: Deskriptive Kennzahlen, Kennzahlen der Tests und die Effektstärke d der Item-Trennschärfe in Abhängigkeit der binären Eigenschaften – nach Abschnitten.

Wenn man sich parallel dazu den Einfluss auf die Item-Trennschärfe getrennt nach Abschnitten ansieht (in Tabelle 14), sieht man trennschärfere Fragen des Typs-A negativ in der Vorklinik (d=-0,24) und mit Aussagen in Vorklinik (d=-0,13) und Klinik (d=-0,14). Der Typ-K zeigt über die Abschnitte hinweg stark schwankende Richtungen und Effektstärken (d -0,22/0,93/-0,17). Beim Fragentyp mit Abbildungen zeigt sich eine bessere Trennschärfe in der Vorklinik (d=-0,22), in der Zwischenklinik wurden nur 5 Items untersucht. Fragen mit Rechnungen sind zwar weniger trennscharf in Vor- und Zwischenklinik (d=0,44/0,71), allerdings wurden auch nur 34 und 6 Items untersucht.

Die teils unterschiedlichen Ergebnisse in den Effektstärken, aber auch in den Richtungen der Effekte sind in Abbildung 29 nochmals grafisch dargestellt.

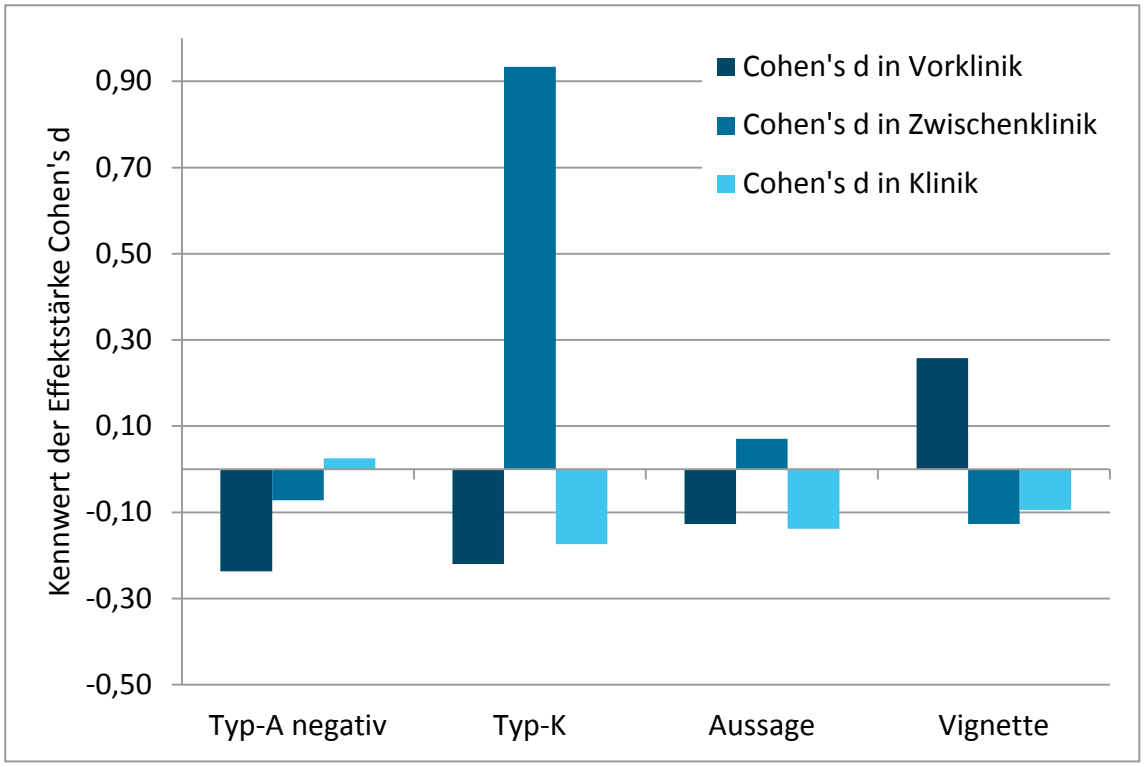

Abbildung 29: Grafische Darstellung der Effektstärken der Item-Trennschärfen in Abhängigkeit der unterschiedlichen Fragetypen und Eigenschaften. Hinweis: Negative Zahlen und Balken nach unten kennzeichnen mehr trennscharfe Fragenitems, positive Zahlen und Balken nach oben weniger trennscharfe Fragenitems.

6.8. Einfluss der Fragetypen und Eigenschaften auf die Distraktoren

In der nachfolgenden Tabelle 15, sowie in den anschließenden Abbildung 30 und Abbildung 31 werden die Eigenschaften von negativen Formulierungen bis zum Vorhandensein von Rechnungen zusammen mit dem Vorkommen von schlechten und sehr schlechten Distraktoren gezeigt.

Der erste Teil der Tabelle zeigt die relative Häufigkeit von wenig oft gewählten Antwortoptionen (von weniger als 5 % der Studierenden gewählt) und wie oft diese in einer Frage vorkommen. Der zweite Teil zeigt parallel dazu die relative Häufigkeit von nie gewählten Antwortoptionen und wie oft diese in einer Frage vorkommen.

NFD-5%	keine <5%	eine <5%	zwei <5%	drei <5%	vier <5%	fünf <5%	p	
Typ-A negativ	0,10	0,12	0,20	0,24	0,34	0,00	0,017	*
Typ-K	0,09	0,15	0,25	0,32	0,20	0,00	0,000	*
Aussage	0,09	0,18	0,24	0,26	0,23	0,00	0,000	*
Vignette	0,03	0,08	0,15	0,34	0,40	0,00	0,000	*
Abbildung	0,06	0,11	0,26	0,33	0,24	0,00	0,011	*
Rechnung	0,07	0,17	0,20	0,20	0,37	0,00	0,681	
Total	371	609	922	1219	1404	5		
NFD-0%	**keine =0**	**eine =0**	**zwei =0**	**drei =0**	**vier =0**	**fünf =0**	**p**	
Typ-A negativ	0,35	0,24	0,20	0,14	0,07		0,068	
Typ-K	0,51	0,26	0,16	0,07	0,01		0,000	*
Aussage	0,49	0,25	0,14	0,08	0,04		0,000	*
Vignette	0,18	0,25	0,26	0,21	0,09		0,000	*
Abbildung	0,54	0,26	0,12	0,06	0,01		0,000	*
Rechnung	0,46	0,32	0,10	0,05	0,07		0,297	
Total	1731	1119	829	568	283			

Tabelle 15: Relative Häufigkeiten, mit der „schlechte“ (NFD-5%) und „sehr schlechte“ (NFD-0%) Distraktoren in Abhängigkeit der Fragetypen und Item-Eigenschaften vorkommen (die Berechnung der Unterschiede wurde mit dem Kruskal-Wallis-Test durchgeführt).

Unter den Ausprägungen, reichend von keiner bis vier Distraktoren wurden von weniger als 5 % der Studierenden gewählt, zeigt meist die Kategorie mit drei oder vier Distraktoren die größte Häufigkeit. Bei den Ausprägungen, die nie gewählten Antwortoptionen betreffend, zeigt meist die Kategorie mit keiner dieser Distraktoren die größte Häufigkeit. Lediglich bei den Vignetten-Fragen ist es die Kategorie mit zwei Distraktoren.

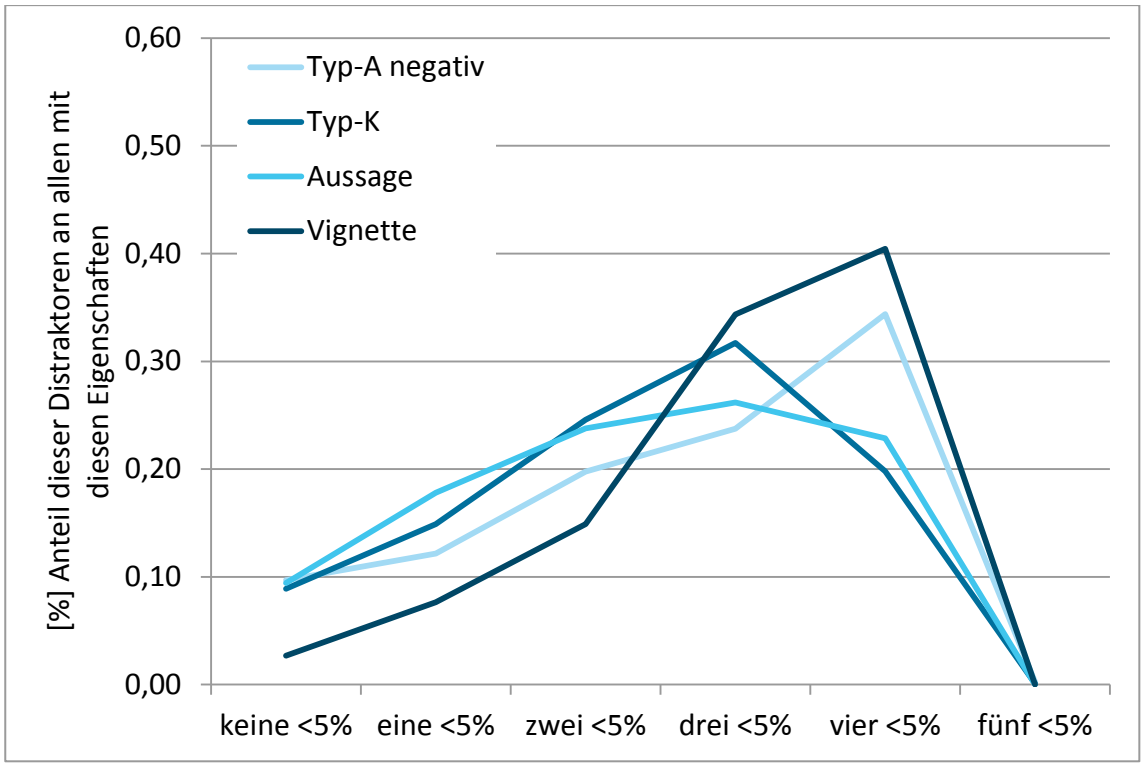

Abbildung 30: Grafische Darstellung der relativen Häufigkeiten, mit der „schlechte“ (NFD-5%) Distraktoren vorkommen.

Leichter verständlich ist es, wenn man sich das dazugehörige Liniendiagramm in Abbildung 30 ansieht: Fragen des Typs-K und Aussagen-basierte Fragen haben in der Regel zwischen zwei und vier schlechte Antwortoptionen, bei negativ formulierten Fragen sind es meistens vier, ebenso wie bei Fragen mit Fallvignetten. Wenn man sich die Häufigkeit des Vorkommens bei den *schlechten* Vignettenfragen ansieht, haben ca. 40 % aller Vignettenfragen vier schlechte Distraktoren – d.h. in der Regel sind dann alle falschen Antwortoptionen schlecht gewählt.

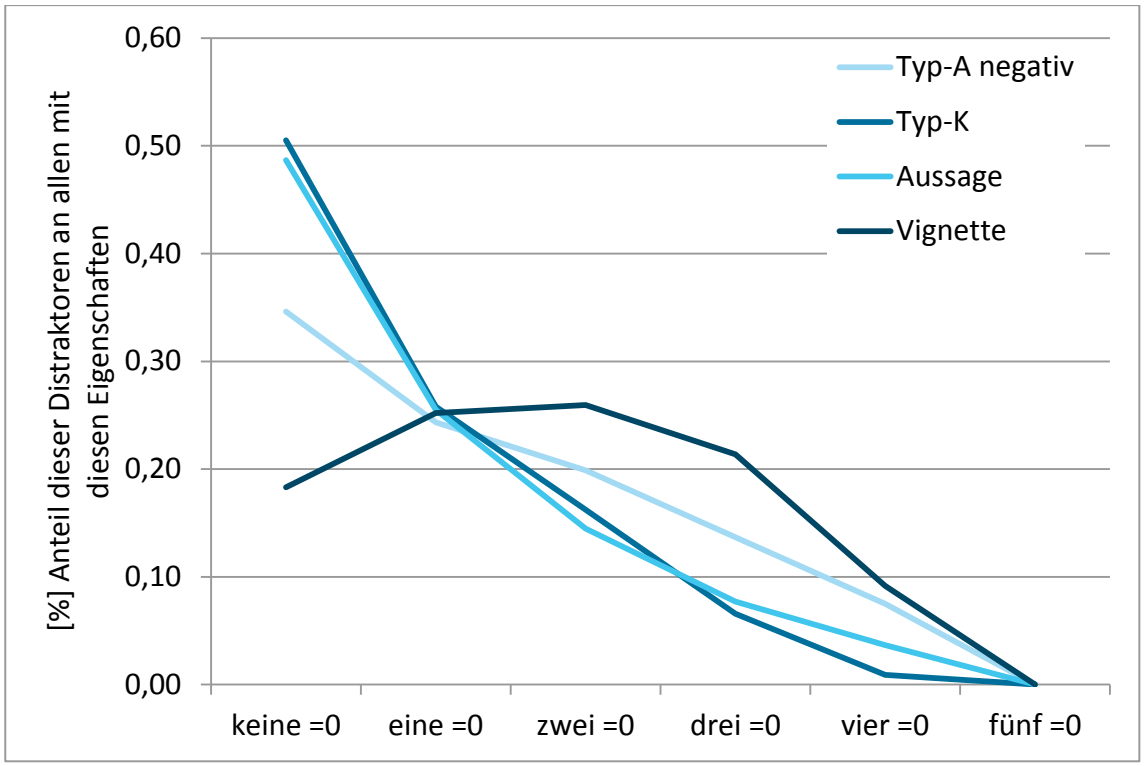

Abbildung 31: Grafische Darstellung der relativen Häufigkeiten, mit der „sehr schlechte“ (NFD-0%) Distraktoren vorkommen.

Auch beim Liniendiagramm zu den nie gewählten Antwortoptionen in Abbildung 31 verhält es sich ähnlich: Fragen des Typs-K und Aussagen-basierte Fragen haben in der Regel wenig sehr schlechte Antwortoptionen, meist ist die Hälfte der Fragen soweit in Ordnung, dass keine sehr schlechte Option vorkommt. Bei den Fragen mit negativen Formulierungen verschlechtert sich das Bild, Fragen mit Vignetten haben auch hier die meisten sehr schlechten Distraktoren. Nicht selten sind zwei oder drei Distraktoren so schlecht, dass keiner der Studierenden sie wählt. Aus einer anderen Perspektive gesehen, sind weniger als 20 % der Vignettenfragen soweit in Ordnung und enthalten keine sehr schlechte und nie gewählte Option.

Analog zur Abbildung 25 werden auch in Abbildung 32 die Kategorien null bis fünf schlechte Antworten pro Fragenitem vereinfacht: auch hier wird der Mittelwert der Kennzahlen *NFD-5%* und *NFD-0%* berechnet und angezeigt, allerdings für alle Fragetypen und Item-Eigenschaften.

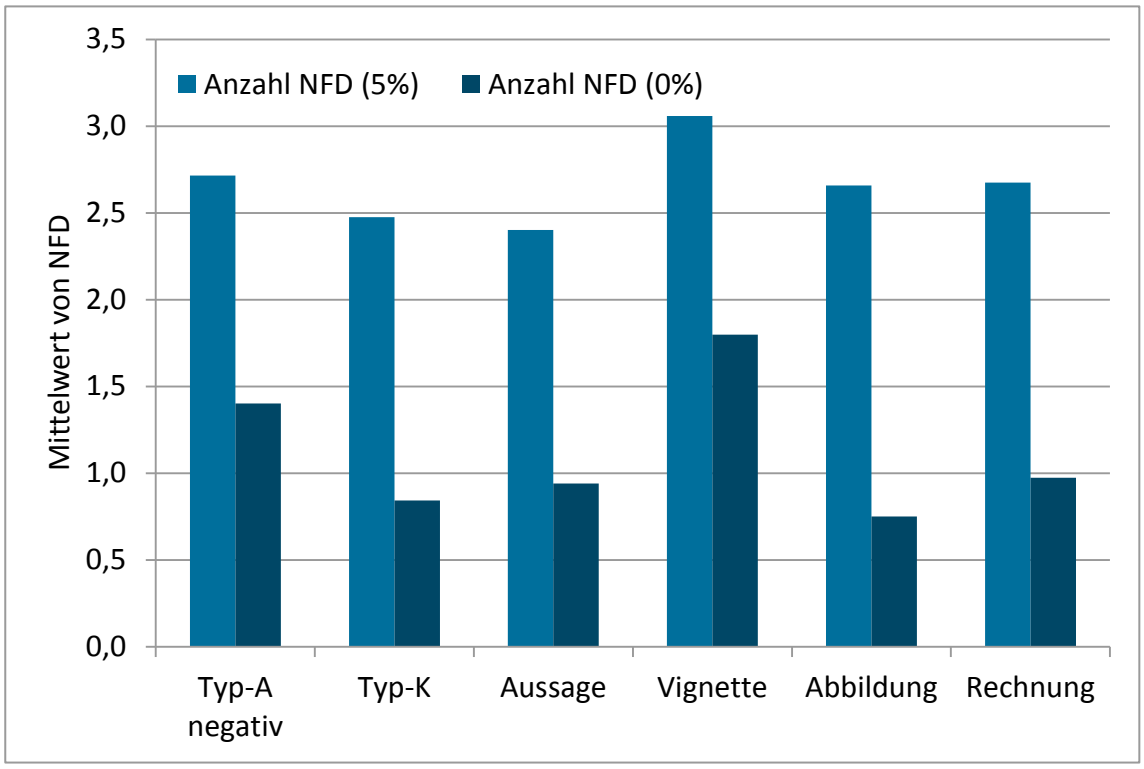

Abbildung 32: Balkendiagramm zur Darstellung der Kennzahlen NFD-5% und NFD-0% in Abhängigkeit von Fragetypen und Item-Eigenschaften.

Der Mittelwert von *NFD-5%* ist mit 3,1 bei Vignetten-Fragen am höchsten, gefolgt von 2,7 bei negativen Fragen und Fragen mit Abbildungen oder Rechnungen. Der Mittelwert von *NFD-0%* ist ebenfalls mit 1,8 bei den Fragen mit Vignetten mit Abstand am höchsten, gefolgt von 1,4 bei negativen Fragen.

6.9. ERGEBNISSE DER PTM-BETEILIGUNG

Wie bereits im Datenteil beschrieben, liegen nur aus den Semestern zwischen dem Wintersemester 2012/2013 und dem Sommersemester 2014 Daten der Medizinischen Universität Graz und den anderen Fakultäten vor. In diesen vier Semestern haben jeweils 406, 328, 639 und 345 Studierende aus Graz teilgenommen. Auf die Jahrgänge in jedem Semester verteilt haben im Mittel 53,7 Studierende (Minimum 7 und Maximum 325) teilgenommen. Die Studierenden des ersten, also frühesten Jahrgangs, waren im dritten Semester des Curriculums an Medizinischen Universität Graz, die des letzten Jahrgangs im 11. Semester des Curriculums.

Alle anderen Fakultäten zusammen haben im Durchschnitt rund 7000 Studierende pro Semester den PTM absolvieren lassen, pro Jahrgang waren es im Durchschnitt 798,2 Studierende.

Die geringste mittlere Punktezahl hatte im WS 2012/2013 der Jahrgang des dritten Semesters mit 11,14 Punkten. Die höchsten mittleren Punktezahlen hatten die Studierenden des 11. Semesters, das trifft auf alle Semester zu, die Punktezahlen lauten: 56,30/72,80/52,79 und 59,34. Im Vergleich dazu war die geringste mittlere Punktezahl der anderen Fakultäten 12,04 (ebenfalls im WS 2012/2013). Ebenfalls vergleichbar waren die höchsten mittleren Punktezahlen im letzten Jahrgang: 74,80/80,95/70,43 und 78,75.

Um einen Überblick über die Leistungen der Studierenden aus Graz und den anderen Fakultäten im Vergleich zu bekommen, wurde für alle Semester und Jahrgänge der Mittelwert über alle mittleren Punktezahlen berechnet und grafisch (in Abbildung 33) dargestellt:

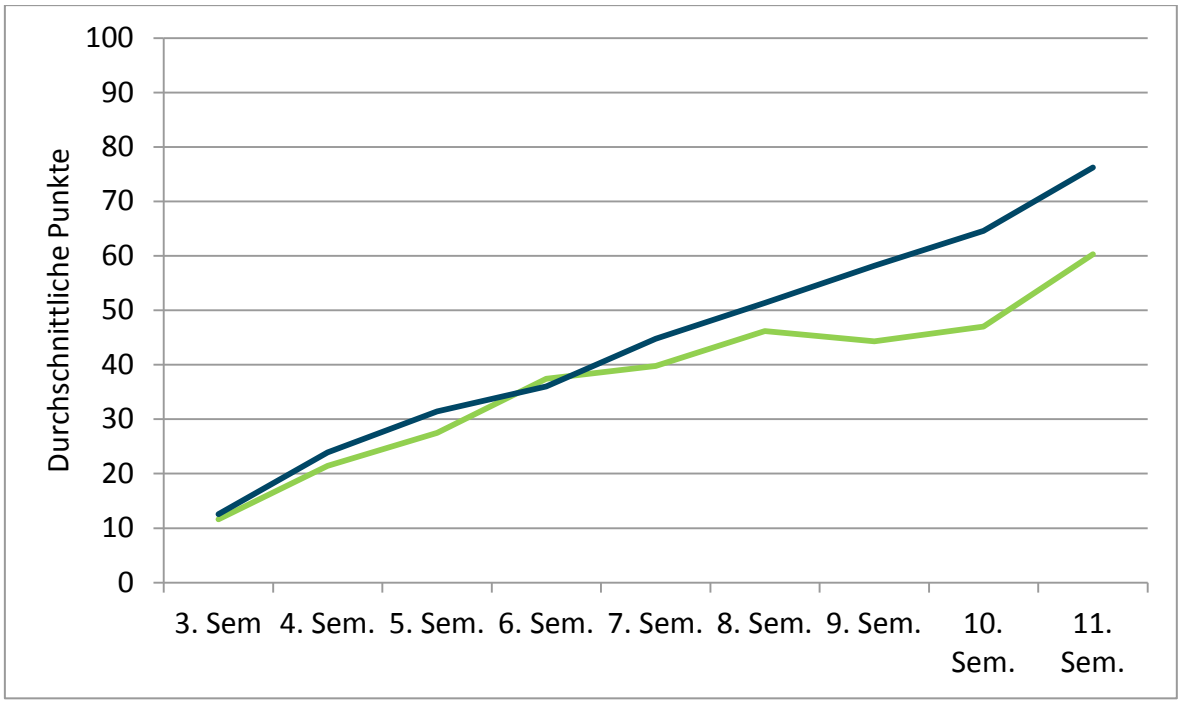

Abbildung 33: Leistungsvergleich zwischen Studierenden der Medizinischen Universität Graz (grün) und den anderen mitteleuropäischen Fakultäten (blau), die Werte wurden wie oben beschrieben gemittelt (auf der x-Achse die Semester, auf der y-Achse die mittleren Punktezahlen).

In der nachfolgenden Abbildung 34 finden sich dieselben grafischen Darstellungen, allerdings separat für jedes Semester der Datenerhebung.

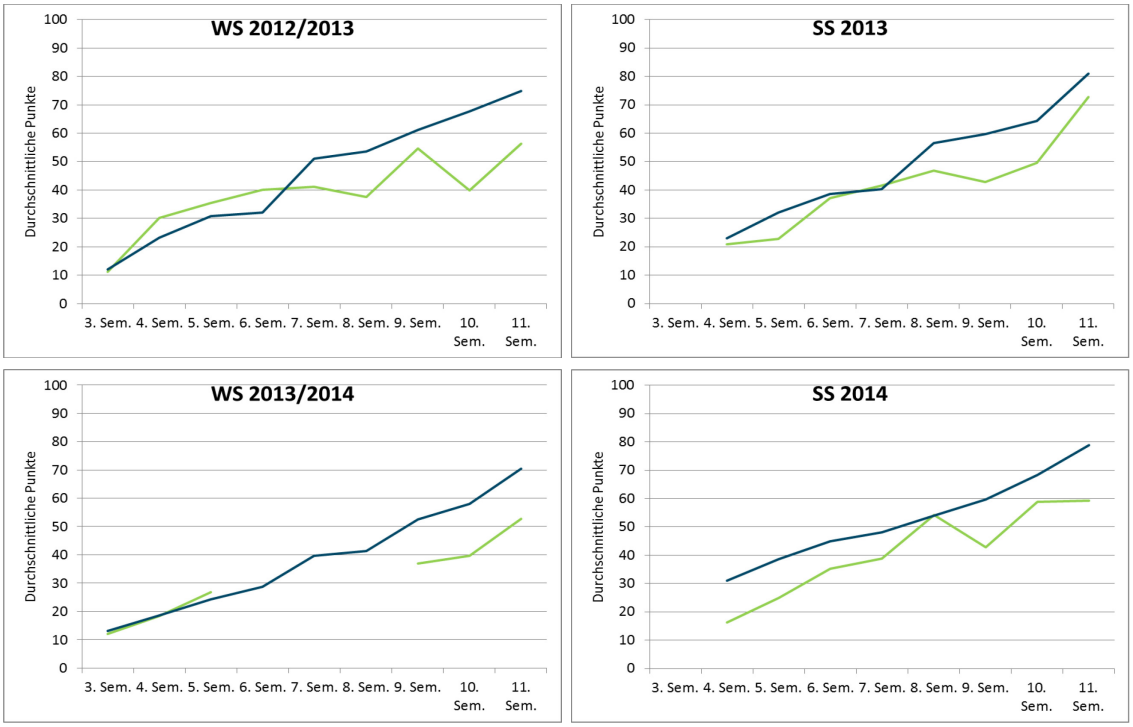

Abbildung 34: Leistungsvergleich, wie in der obigen Übersicht, allerdings separat für jedes Semester (Achsenbeschriftung wie oben).

Diese vier Abbildungen zeigen ein sehr ähnliches Bild. Im Bereich der Vorklinik schwanken die Leistungen teils deutlicher: Die Leistung im WS 2012/2013 ist eher überdurchschnittlich, im SS 2014 jedoch eher unterdurchschnittlich. Abgesehen von fehlenden Werten (WS 2013/2014) und Schwankungen ergibt sich jedoch im klinischen Abschnitt ein relativ gleichbleibendes Bild.

In der darauf folgenden Abbildung 35 ist derselbe grafische Aufbau zu finden, allerdings stellen hier die beiden Graphen unterschiedliche Studienjahre innerhalb der Medizinischen Universität Graz dar.

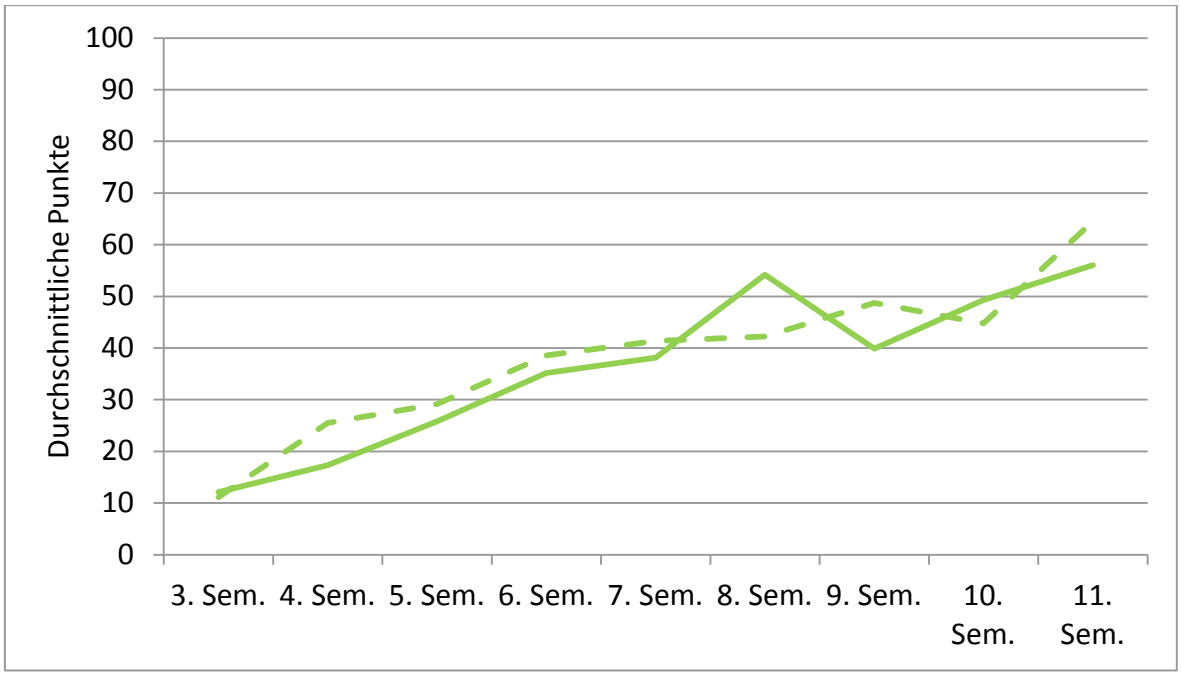

Abbildung 35: Leistungsvergleich zwischen Studierenden der Medizinischen Universität Graz aus zwei unterschiedlichen Studienjahren, grün-strichliert das Studienjahr 2012/2013 und grün-durchgehend das Studienjahr 2013/2014 (auf der x-Achse die Semester, auf der y-Achse die mittleren Punktezahlen).

In dieser Abbildung werden die ersten beiden Semester (2012 und 2013) mit den zweiten beiden Semestern (2013 und 2014) verglichen. Dabei zeigen sich eine konstante Verschlechterung im Bereich der Vorklinik und wechselnde Entwicklungsrichtungen im Bereich der Klinik. Vor allem im 8. Semester ist eine deutliche Veränderung sichtbar, die möglicherweise durch das wichtige, weil internistische Modul 16 verursacht wurde.

Ergänzend und den Ergebnisteil abschließend sind in der Tabelle 16 alle PTM-Ergebnisse, die für den Vergleich herangezogen wurden, aufgelistet:

Semester	Jahrgang	MUG MW	MUG SD	MUG N	And MW	And SD	And N	ES
WS 12/13	3	11,14	9,57	49	12,04	13,22	901	-0,07
	4	30,17	14,17	12	23,23	15,41	507	0,45
	5	35,42	19,01	102	30,74	18,39	1077	0,25
	6	40,00	18,51	10	32,12	23,09	559	0,34
	7	41,18	20,72	114	50,98	26,92	1055	-0,37
	8	37,50	19,90	10	53,58	25,27	533	-0,64
	9	54,58	24,15	79	61,17	27,52	1058	-0,24
	10	39,90	23,74	10	67,76	29,57	700	-0,95
	11	56,30	28,96	20	74,80	31,94	337	-0,58
SS 13	3				12,96	12,38	483	
	4	20,85	12,42	67	22,91	16,88	1302	-0,12
	5	22,89	21,26	9	32,03	19,39	698	-0,47
	6	37,12	20,48	75	38,51	21,55	1414	-0,06
	7	41,57	20,97	7	40,35	22,59	504	0,05
	8	46,90	21,18	81	56,52	25,98	1032	-0,38
	9	42,90	19,38	10	59,57	27,37	512	-0,61
	10	49,65	25,21	69	64,31	29,23	1310	-0,50
	11	72,80	20,64	10	80,95	28,04	294	-0,29
WS 13/14	3	12,14	12,42	325	13,11	14,01	1011	-0,07
	4	18,33	19,92	9	18,49	16,00	514	-0,01
	5	26,83	19,11	113	24,20	17,40	1119	0,15
	6				28,71	20,76	640	
	7	37,42	23,11	73	39,69	24,25	1065	-0,09
	8				41,43	26,49	528	
	9	36,93	20,59	73	52,48	27,24	902	-0,58
	10	39,71	19,56	17	58,07	29,33	655	-0,63
	11	52,79	21,16	29	70,43	25,31	218	-0,71
SS 14	3				20,68	14,25	495	
	4	16,33	14,33	21	30,97	17,04	1665	-0,86
	5	24,85	21,06	13	38,68	21,00	592	-0,66
	6	35,14	18,24	86	44,83	21,98	1505	-0,44
	7	38,86	24,85	7	48,00	23,06	587	-0,40
	8	54,16	22,04	101	53,90	26,88	1069	0,01
	9	42,79	15,97	14	59,62	25,65	521	-0,66
	10	58,87	22,20	71	68,32	28,66	1242	-0,33
	11	59,34	26,82	32	78,75	29,23	130	-0,68

Tabelle 16: Übersicht aller PTM Ergebnisse für die Med Uni Graz und der Vergleichsgruppe der anderen Universitäten über vier Semester, beginnend beim Wintersemester 2012/2013.

7. Diskussion

7.1. Kurzüberblick

7.1.1. Vorkommen und Verteilung

Beim Überblick ist festzuhalten, dass die meisten Items aus dem klinischen Abschnitt untersucht wurden, da dort viele kleine Prüfungen absolviert wurden. Jeder Prüfungsdurchgang mit mehr als 25 Studierenden wurde in die Auswertung aufgenommen. Bei weniger Studierenden wären weder das Vorkommen und die Verteilungsmuster der Fragetypen und Item-Eigenschaften, noch die errechneten Kennwerte Item-Schwierigkeit und Item-Trennschärfe ausreichend aussagekräftig oder belastbar.

Jede fünfte Frage ist entweder negativ formuliert, eine Typ-K Frage oder eine Aussagen-basierte Frage (Vorkommen der drei Eigenschaften jeweils rund 20 %). Fragen mit Fall-Vignette kommen nur in rund 6 % vor. Bei den Abbildungen verhält es sich ebenso, Rechnungen kommen hingegen noch seltener vor.

Über beide Jahre gerechnet kann man sagen: Negative Fragen sind mit knapp 30 % in der Klinik häufiger, Typ-K Fragen mit gut 50 % in der Vorklinik. Bei Aussagen-basierten Fragen gibt es keine so schiefe Verteilung, am häufigsten kommen sie in der Vorklinik vor, am seltensten in der Zwischenklinik. Vignettenfragen sind wiederum in der Klinik am häufigsten, absolut gesehen jedoch nur zu 10 %. Fragen mit Abbildungen oder Rechnungen kommen wieder in der Vorklinik häufiger vor, insgesamt aber auch sehr selten (2-5 %).

Die Kombination an Fragetypen und Item-Eigenschaften betreffend, kommen Abbildungen vermehrt in Fragen des Typs-K oder in Aussagen-basierten Fragen vor. Ebenfalls gehäuft kommen Aussagen im Zusammenhang mit Typ-K oder negativen Formulierungen vor. Vignetten-Fragen sind gelegentlich noch negativ formuliert, kommen jedoch nie in Kombination mit Fragen des Typs-K, Aussagen oder Abbildungen vor.

7.1.2. Schwierigkeit und Trennschärfe

Unabhängig der Wahl des Lagemaßes (Mittelwert oder Median) sind die Prüfungen der Klinik am leichtesten (Median 0,875) und die Prüfungen im Bereich der Vorklinik am meisten trennscharf (Median 0,290). Die Verteilung des Item-Schwierigkeitsgrades an Hand eines Histogramms zeigt die extreme Einseitigkeit, also das Dominieren der leichten Fragenitems. In der Klinik fallen knapp 2/3 der Fragen in die Kategorie Item-Schwierigkeit >0,8. Bezogen auf Fragetypen und Item-Eigenschaften haben besonders Vignettenfragen einen hohen Anteil an Fragen mit einer Item-Schwierigkeit >0,8 (72 %). Aber auch negative Fragen haben einen hohen Anteil an dieser Kategorie (57 %).

7.1.3. DiDo und die Distraktoren

Der Anteil an Fragen, bei denen einer der Distraktoren häufiger gewählt wurde als die korrekten Antworten, liegt zwischen 2 und 10 % im jeweiligen Jahr und Abschnitt. Am häufigsten kommen solche Fragen in der Zwischenklinik vor, gefolgt vom klinischen Abschnitt.

In der Klinik haben die Distraktoren deutlich schlechtere Qualität – sowohl, wenn man sich die Anzahl an Distraktoren, welche von weniger als 5 % der Studierenden gewählt wurden, wie auch die, die nie gewählt wurden, ansieht. Auch die absoluten Zahlen sind hierbei eindeutig: Knapp 40 % der Klinikfragen haben vier schlechte Distraktoren. Bei den sehr schlechten, nie gewählten Distraktoren, sind nur knapp 25 % aller Klinikfragen fehlerfrei. Immerhin noch 10 % haben vier sehr schlechte Distraktoren.

Der Anteil an schlechten Distraktoren pro Fragenitem (NFD-5%) ist im klinischen Bereich am Höchsten mit 2,9. Das bedeutet auch, dass nur mehr ein einziger *besserer* Distraktor überbleibt, womit die Ratewahrscheinlichkeit bei 50 % liegt. Der Anteil an sehr schlechten Distraktoren (NFD-0%) in der Klinik liegt auch noch bei 1,7. Im Durchschnitt sind also knapp zwei Distraktoren nie gewählt.

7.1.4. EINFLUSS DER FRAGETYPEN

Zu Beginn wurde die Verteilung der einzelnen Fragetypen und Item-Eigenschaften auf die einzelnen Abschnitte erläutert, danach wurden die beiden Kennwerte Item-Schwierigkeit und Item-Trennschärfe in den Abschnitten zusammengefasst. Der nächste logische Schritt ist nun, zu überprüfen, ob die Verteilung der Eigenschaften, deren Auswirkungen auf die Kennwerte und die tatsächlich beobachteten Kennwerte zusammenpassen.

Der Fragentyp-K ist signifikant schwerer (d=0,162), ebenso der Aussagen-basierte Fragentyp (d=0,231). Der Fragentyp, bei dem Fall-Vignetten zum Einsatz kommen, ist dagegen signifikant leichter (d-0,353). Parallel dazu ist der Fragentyp-A negativ signifikant weniger trennscharf (d=0,088), die Fragetypen-K (d=-0,254), Aussagen-Fragen (d=-0,165) und Abbildungsfragen (d=-0,492) dagegen deutlich und signifikant trennschärfer.

Wenn man nun die Verteilung der Fragetypen und Item-Eigenschaften auf die einzelnen Abschnitte und deren Auswirkungen zusammen ansieht, ergibt sich folgendes Bild: Fragen des Typs-K und Aussagen-basierte Fragen kommen vor allem im Bereich der Vorklinik vor und heben die Item-Schwierigkeit an, d.h. machen die Prüfungsfragen schwerer lösbar. Negative Fragen und Fragen mit Vignette kommen häufiger im Bereich der Klinik vor und machen die Prüfungen weniger trennscharf (negative Fragen) und auch leichter (Vignetten).

Dasselbe gilt für die Distraktoren: Bestimmte Fragetypen haben spezifische Eigenschaften der Distraktoren, und da diese Typen eine bestimmte Verteilung über die Abschnitte haben, verteilen sich auch die Distraktoren-Eigenschaften entsprechend. Der Fragentyp-K und Aussagenfragen, welche vor allem in der Vorklinik vorkommen, haben relativ gute Distraktoren (in absoluten Zahlen auch verbesserungswürdig). Im Vergleich dazu haben Fragen des Typs-A negativ und Vignettenfragen, welche vor allem im Bereich der Klinik vorkommen sehr schlechte und somit stark verbesserungswürdige Distraktoren.

Dieselbe Verteilung und dieselben Aussagen über die Distraktoren gelten für schlechte und sehr schlechte Distraktoren, also unabhängig, ob von weniger als 5 % der Studierenden oder nie gewählt.

7.1.5. ERGEBNISSE DES PTM

Kurz zusammengefasst zeigt sich hier ein *Einbruch* der Studierendenleistung ab dem frühen klinischen Abschnitt (7. Semester) und damit eingeleitet, ein *Zurückbleiben* der eigenen Studierenden im klinischen Bereich. Vor allem im 9. Semester wird der Abstand größer, im 11. Semester geringfügig kleiner. Auch in der Semester-Einzelauswertung zeigt sich dieses Phänomen in Form eines konstanten Bildes. Wenn man die ersten beiden Semester mit den letzten beiden Semester vergleicht, könnte man von punktuellen Verbesserungen sprechen, am deutlichsten im 8. Semester.

7.2. LITERATURVERGLEICH UND INTERPRETATION

Der Überblick über alle untersuchten Einheiten muss nicht explizit kommentiert und interpretiert werden, da diese Kennzahlen nur die administrative Lage wiederspiegeln. Wie viele Prüfungsdurchgänge es gegeben hat, wie viele Items dabei im Einsatz waren und wie viele Studierende angetreten sind, haben keinen Einfluss auf Item-Eigenschaften und deren Auswirkungen auf Kennwerte wie beispielsweise die Item-Schwierigkeit. Die Zahlen sind in beiden Studienjahren 2011/2012 und 2013/2014 ähnlich und damit kann man auch von relativ stabilen Rahmenbedingungen ausgehen.

7.2.1. VORKOMMEN UND VERTEILUNG

Das Vorkommen einzelner Fragetypen und Item-Eigenschaften und die Verteilung derselben sind teilweise gut nachvollziehbar und gut zu begründen. Teilweise spiegelt die Verteilung jedoch auch eine spezielle Situation wider.

19 % der Fragen im untersuchten Fragenpool sind *Typ-A negative Fragen*, in der Klinik kommen sie am häufigsten mit 27 % vor. Obwohl ein Drittel der untersuchten Fragen negativ formuliert sind, wird dieser Fragentyp nicht von der Literatur empfohlen. Bei Haladyna, aber auch bei Rodriguez ist die Zustimmung zur Regel der Vermeidung negativer Formulierungen groß (Haladyna & Downing 2002; Rodriguez 1997). Bei den Auswirkungen dieses Fragentyps ist sich die Literatur wenig einig, die Bandbreite reicht von weniger schwierigeren bis schwierigeren Typ-A negativen Fragen. Einige Studien, die in den beiden Arbeiten erwähnt werden, sprechen sich für keinen Unterschied zwischen positiv und negativ formulierten Fragen aus. Rodriguez spricht von einem uneinheitlichen Bild und auch Krebs meint, der Schwierigkeitsgrad und die Item-Trennschärfe ist mit denen der positiven Typ-A Frage vergleichbar (Krebs 2008).

Negative Typ-A Fragen können in speziellen Fällen eingesetzt werden, nämlich wenn es um das Kennen und damit Prüfen von Ausnahmen geht. Meist werden sie jedoch aus anderen Gründen eingesetzt. Sie sind in der Praxis leichter zu erstellen, wie in Kapitel 4.2.3 erläutert wurde. Warum aber kommen Sie deutlich öfter im klinischen Bereich vor? Möglichweise ist das deshalb der Fall, da hier das Fragenerstellen schwerer und ein Ausweichen in negative Formulierungen dringlicher ist.

Der *Typ-K*, der ebenfalls 20 % der Fragen im untersuchten Fragenpool ausmacht, ist ein Fragentyp, der ebenfalls nicht mehr empfohlen wird, aber trotzdem stellenweise noch gerne eingesetzt wird. Im Bereich der Vorklinik ist jede zweite Frage (52 %) dieses Typs. Auch hier sprechen sich Haladyna und Rodriguez gegen die Verwendung aus (Haladyna & Downing 2002; Rodriguez 1997).

Es handelt sich um einen *schlechten* Fragentyp, der durch Fragen des MTF-Typs oder auch Kprim genannt, ersetzt werden sollte. In der Literatur, die von Rodriguez beschrieben wurde, zeigt sich der erhöhte Schwierigkeitsgrad dieses Fragentyps in sechs von sieben Arbeiten. Er sieht abschließend einen erhöhten Schwierigkeitsgrad, der erwünscht oder auch nicht erwünscht sein kann, aber auch eine geringere Trennschärfe dieser Fragen. Albanese sieht im Gegenzug einen leichteren Fragentyp, ähnlich wie Krebs (Albanese 1993; Krebs 2004). Letzterer beschreibt eine 50 % Ratewahrscheinlichkeit, sobald man eine einzige Aussage als eindeutig richtig oder eindeutig falsch identifiziert hat.

Viele Lehrende verweisen darauf, dass in dem Format mehr als eine Aussage richtig sein kann. Sie sprechen oft auch die Komplexität mit dem zweistufigen Lösungsprozess an und finden das Erschweren einer Frage auf diese Weise sinnvoll. Das vermehrte Vorkommen in naturwissenschaftlichen Grundlagen-Fächern anstatt in klinischen Fächern verwundert zumindest nicht. Wenn man den Aufbau mit vier zu bewertenden Aussagen sieht, liegt das Verwenden klarer naturwissenschaftlicher Fakten und Aussagen nahe. Durch diese angesprochene Komplexität und diesen zweistufigen Denkprozess ist es wenig verwunderlich, dass dieser Fragentyp zu schwereren Fragen führt.

Ein *Aussagen-basierter Typ* liegt im Durchschnitt bei 19 % aller Fragen vor. Dieser Typ kommt in allen Abschnitten vor, am häufigsten in der Vorklinik mit 31 %, etwas weniger in der Klinik mit 17 % und am seltensten in der Zwischenklinik mit 9 %. Diese Art von Frage ist auch deutlich leichter zu erstellen, da sie keinen Fokus besitzen muss. Die *breite* Frage lautet: „Welche Aussage zum Thema *xy* ist richtig?“ Unter eine solche Frage kann man sehr viele verschiedene, auch nicht oder nur wenig zusammenpassende, Aussagen schreiben. Der Kontext ist hierbei ein sehr weiter. Da der Fokus nicht gegeben sein muss, ist das Erstellen oder besser Zusammenstellen der Aussagen leichter.

Die Fragen selbst sind dann aber auch etwas schwerer, da man Aussagen sehr unterschiedlicher Dimensionen miteinander vergleichen muss (diese Thematik wurde ebenso in Kapitel 4.2.3. beschrieben). In der Literatur findet sich wenig Explizites, auch wenn der Typ unter dem Begriff *statement type* gelegentlich vorkommt.

Fragen mit Vignetten oder kurzen Fallbeschreibungen kommen nur in 6 % vor, mit Abstand am häufigsten in der Klinik mit 10 %. Dabei liegt das Verwenden klinischer Fallvignetten im dritten und klinischen Abschnitt klar auf der Hand. Man beschäftigt sich nicht mehr mit naturwissenschaftlichen und isolierten Fakten, sondern versucht, diese zusammen mit anderen Elementen in der Problemlösung einzusetzen.

Diese Szenarien sind mehrdimensionaler und komplexer, die Lösung oft weniger eindeutig. Um dann Diskussionen um die richtige Antwort zu vermeiden, entschließt man sich offenbar oft für sehr eindeutige Antwortoptionen, d. h. Optionen, die eindeutig richtig oder falsch sind. Damit werden die Fragen auch leichter, wie man es bei den Vignettenfragen im klinischen Abschnitt sehen kann, und die Anzahl an schlechten und sehr schlechten Distraktoren ist vergleichsweise hoch.

Was die Auswirkungen dieses Fragentyps betrifft sind sich die Expertinnen und Experten nicht einig. Caballero spricht von einem schwierigen Typ, Ikah von einem leichteren und Phipps sieht den Typ gleich schwer wie Nicht-Vignetten (Caballero et al. 2014; Ikah et al. 2015; Phipps & Brackbill 2009). Im Detail hat Caballero 516 Items untersucht und neben dem erhöhten Schwierigkeitsgrad auch eine erhöhte Trennschärfe gesehen. Ikah spricht von leichteren, aber auch mehr trennscharfe Fragen, hat aber nur eine kleine Stichprobe ausgewertet. Phipps sieht gleich schwere Fragenitems, zudem mit niedrigerer Trennschärfe.

Fragen mit Abbildungen kommen bei 5 % der Fragen vor, in der Vorklinik bei 18 % der Fragen. Sie sind geringfügig leichter, aber deutlich trennschärfer als Fragen ohne Abbildungen. Andererseits hat Holtzmann im Jahr 2009 anhand der USMLE-Prüfungen gezeigt, dass mit Multimedia angereicherte Fragenitems schwerer sind und weniger trennscharf (Holtzman et al. 2009).

Der Umstand, dass klinische Abbildungen beinahe nicht (nur in einem einzigen Fragenitem!) in klinischen Fragen eingesetzt werden, enttäuscht ein wenig. Ob Fragen zu den Bildrechten und datenschutzrechtliche Fragen als Ursache dafür gesehen werden können, darüber kann nur spekuliert werden.

Die Kombination von bestimmten Fragetypen und Item-Eigenschaften betreffend, gibt es wenige Überraschungen. Die im ersten Abschnitt vorherrschenden Eigenschaften Typ-K, Aussagen und Abbildungen kommen häufiger in Kombination vor. Ein Grund ist natürlich im gemeinsamen Vorkommen zu sehen.

Vignetten kommen hingegen nie zusammen mit Typ-K Fragen, Aussagen oder Abbildungen vor: Ein Grund ist wieder in der unterschiedlichen Verteilung über die drei Abschnitte hinweg zu sehen, wobei doch verwundert, dass klinische Fallvignetten nie eine klinische Abbildung – beispielsweise ein Foto einer Hautefflo-reszenz oder einer sonstigen typischen Blickdiagnose – enthalten. Weder Abbildungen, noch Vignetten oder Fragen des Typs-K kommen häufig in negativ formulierten Fragen vor.

Negative Formulierungen kommen dagegen häufig zusammen mit Aussagen vor. Dieses Ergebnis überrascht nicht, da die bereits genannte Formulierung „welche Aussage trifft zu“ leicht in „welche Aussage trifft NICHT zu“ abgeändert werden kann. Die Möglichkeit des Formulierens von negativen Varianten bei Fragen des Aussagen-Typs ist damit offensichtlich.

7.2.2. SCHWIERIGKEIT UND TRENNSCHÄRFE

Der Mittelwert der Item-Schwierigkeit ist über alle Abschnitte und Jahre gesehen 0,770 und der Median ist 0,833. Schwankungen zwischen den beiden untersuchten Jahren sind gering. Was man an den beiden Lagemaßen Mittelwert und Median bereits ablesen kann, bestätigt sich auch im Histogramm der Item-Schwierigkeit: eine deutlich linksschiefe Verteilung. Der Anteil an sehr leichten Fragen mit einer Schwierigkeit über 0,8 liegt in Summe bei 56 %. Dagegen hat die mittlere Kategorie (0,2 bis 0,8) nur 42 % und die schwerste Kategorie (<0,2) nur 2 % Anteil. Ein Lagemaß um einen Wert von 0,5 (eventuell bis 0,6 – siehe Möltner) wäre wünschenswert, wird aber deutlich verfehlt.

Möltner empfiehlt eine Item-Schwierigkeit zwischen 0,4 und 0,8, als Begründung für die Asymmetrie gibt er die Motivation der Teilnehmenden an (Möltner et al. 2006).

Bei der Item-Trennschärfe ist die Situation ausgeglichener: Der Mittelwert über alle Abschnitte und Jahre hinweg ist 0,239 und der Median 0,243. Das Histogramm zeigt auch eine annähernde Normalverteilung, mit einer guten Verteilung der Werte. Lediglich beim Wert 0 gibt es eine Häufung, was auf die vielen Fragen mit einer Schwierigkeit von 0 bzw. vor allem 1 zurückzuführen ist. Analog zur Schwierigkeit beschreibt Möltner auch die Zielbereiche der Trennschärfe. Werte >0,3 wären wünschenswert und gut, Werte zwischen 0,2 und 0,3 sind akzeptabel und Werte zwischen 0,1 und 0,2 nur marginal.

So hätte der untersuchte Fragepool immerhin akzeptable Durchschnittswerte, wobei hier die Variabilität zwischen den beiden Jahren größer ist. Zusätzlich ist auch noch anzumerken, dass trotz passender Durchschnittswerte, ein nicht unerheblicher Anteil einen marginalen und sogar negativen Trennschärfe-Wert hat.

Auf Abschnitte bezogen, ist der klinische der leichteste, wenn auch die Vorklinik nicht viel schwerer ist. Der Bereich der Zwischenklinik ist, relativ zu den anderen beiden, am schwersten. Das bedeutet, dass in der Klinik im Durchschnitt 80 % der Studierenden alle Fragenitems dieses Abschnitts auf Anhieb korrekt beantworten (im Vergleich dazu liegt der Median sogar bei 0,875). Diese letzte Zahl bedeutet, dass die Hälfte der Fragen eine Item-Schwierigkeit über 0,875 oder gerundet 0,9 hat – also von mehr als 90 % der Teilnehmenden korrekt beantwortet wird. Wenn man die Schwierigkeit in Kategorien ausdrücken möchte, so fallen 63 % der Klinik, aber nur 52 % der Vorklinik und 40 % der Zwischenklinik in die Kategorie über 0,8. Die Kategorie Schwierigkeit unter 0,2 ist beinahe nicht existent (maximal 3 %). Die Trennschärfe sinkt kontinuierlich vom ersten zum dritten Abschnitt, von 0,297 über 0,265 zu 0,200 (im Mittel). Beides sind Zeichen, dass das Prüfen im klinischen Kontext mit mehr Unschärfe verbunden ist.

Das deutliche Überwiegen zu leichter Fragen kann mehrere Ursachen haben, wobei das fehlende Schreiben von solchen Fragen sicher vorrangig zu nennen ist. Schwere Fragen – wenn auch nicht trickreiche Fragen – sind auch schwer zu erstellen, benötigen Schulung, Training und Gegenlesen. Es ist leichter und geht schneller, wenn man einfache Sachverhalte prüft, es ist ebenso leichter und müheloser, offensichtliche Distraktoren nieder zu schreiben. Wenn man ungewollte Lösungshinweise einbaut, weil man sich der Thematik nicht bewusst ist, werden Items leichter. Andererseits muss man mehrdeutig formulieren und bewusst trickreich agieren, damit man Fragenitems erschwert.

Natürlich kommen auch andere Faktoren ins Spiel: Wenn man beispielsweise wenig neue Fragen erstellt hat, unter den Studierenden Fragen bereits gut bekannt sind, kommt es auch zu einer Verschiebung in Richtung geringer Item-Schwierigkeit.

Die Item-Trennschärfe ist vor allem beim Curriculum der Medizinischen Universität Graz nur bedingt aussagekräftig, da die Module von unterschiedlichen Kliniken und Instituten zusammen veranstaltet werden und daher eine große natürliche Heterogenität vorhanden ist. Diesen Kennwert kann man daher nur grob interpretieren: Allerdings fällt auch hier auf, dass Fragenitems der Klinik schlechter abschneiden. Möglicherweise können auch hier die komplexen Sachverhalte nicht so präzise formuliert und geprüft werden.

7.2.3. Die Distraktoren und DiDo

Die Ergebnisse zum Schwierigkeitsgrad und der Qualität der Distraktoren sind in enger Beziehung zu sehen. Schlechte Antwortoptionen und damit Distraktoren sind Ursachen für zu leichte und auch zu wenig trennscharfe Fragenitems. Bei 8 % der Fragen sind alle Antwortoptionen gut und brauchbar, bei 13 % ist ein Distraktor nicht funktionierend, bei 20 % zwei, bei 27 % drei und bei 31 % alle vier Distraktoren. In diesem Zusammenhang bedeutet *nicht funktionierend*, dass sie von weniger als 5 % der Studierenden gewählt wurden.

Bei sehr schlechten Distraktoren, die nie gewählt wurden, ist das Vorkommen geringer, aber auch relativ hoch. In 38 % sind keine sehr schlechten Distraktoren vorhanden, in 18 % zwei und vier sehr schlechte Distraktoren immer noch in 6 %. Diese Zahlen verschlechtern sich in der Klinik deutlich, sowohl bei den selten, wie auch bei den nie gewählten Distraktoren.

Warum die verwendeten Distraktoren gerade im klinischen Bereich so viel Raum für Verbesserungen lassen, darüber können auf Grund der Datenlage nur Vermutungen angestellt werden.

Es ist anzunehmen, dass die realitätsnäheren und komplexeren Themen der Klinik im Vergleich zu den naturwissenschaftlichen Themen der Vorklinik schwerer in gute Fragen zu fassen sind. Im Bereich der Vorklinik dominieren beispielsweise klare physikalische, chemische oder physiologische Sachverhalte. Aufgrund der klareren Trennung zwischen korrekt und inkorrekt ist auch das Erstellen eindeutiger Fragen einfacher, ohne dass die Distraktoren zu offensichtlich und plump falsch erscheinen.

Klinische Sachverhalte sind mehrdimensionaler und komplexer, benötigen oft ein probabilistisches Denken und sind in Summe auch strittiger. Um aber dann die Anzahl der Diskussionen über solche Items so gering wie möglich zu halten, werden die Fragen mit Hilfe plakativer Antwortoptionen vereinfacht. Damit werden Sie leichter und auch weniger trennscharf.

Wenn man sich die Arbeit von Tarrant ansieht, bemerkt man ein ähnliches Bild (Tarrant et al. 2009): Die Fragenitems haben im Durchschnitt nur 1,54 (SD 0,88) funktionierende Distraktoren, dabei sollten sie doppelt so viele haben. Es wurden nämlich Fragen mit drei Distraktoren untersucht, allerdings haben nur 14 % aller Items drei funktionierende Distraktoren. In Summe sind auch nur 52 % aller Distraktoren funktionierend und effektiv. Ihre Schlussfolgerung ist einfach: Sie vermutet, dass es für Lehrende schwierig ist, plausible Distraktoren zu generieren.

Wenn man die Darstellung der Distraktoren-Qualität von Tarrant nimmt und für den eigenen analysierten Fragenpool ansieht, ist NFD-5% im Bereich der Klinik knapp drei, was bedeutet, dass drei Distraktoren sehr selten gewählt werden und daher nur mehr ein Distraktor von der korrekten Antwortoption ablenken kann. Solche Fragen gleichen dann Fragen mit 50 % Ratewahrscheinlichkeit und können problemlos mit dem Münzwurf in Verbindung gebracht werden.

Aber auch in den beiden anderen Abschnitten ist NFD-5% nicht viel niedriger: in der Vorklinik 2,5 und in der Zwischenklinik 2,3. NFD-0% ist in der Klinik 1,7 und damit deutlich kleiner als NFD-5%, aber die Zahl bedeutet auch, dass diese Distraktoren von keinem der Studierenden gewählt wurde. NFD-0% ist in der Vorklinik und in der Zwischenklinik mit 0,8 weniger als halb so groß. Bei diesen Analysen wurden aber nur Prüfungsdurchgänge herangezogen, bei denen mindestens 25 und im Durchschnitt sogar 100 Studierende antreten – was die Aussagekraft über diese zwei Kennwerte noch verschärft.

Ein ähnlicher Parameter, der für die Qualität von Fragenitems stehen kann, ist der Anteil an Fragen, bei denen Distraktoren häufiger gewählt werden als die korrekten Antwortoptionen. Der DiDo-Anteil ist in Summe immerhin bei 5,5 % und erreicht Spitzenwerte von 10,3 %. So hoch ist dann der Anteil an Fragen, die so gestellt sind, dass die Studierenden *systematisch* die Frage falsch interpretieren und beantworten. Man könnte auch pointiert formulieren, dass dieser Wert proportional zum Schulungsbedarf der Autorinnen und Autoren ist.

7.2.4. Einfluss der Fragetypen auf Schwierigkeit/Trennschärfe

Teilweise wurden gewisse Erklärungsmodelle bereits angedeutet, nun jedoch noch einmal systematisch und mit statistischen Tests verifiziert: *Negative Fragen* haben keinen signifikanten Einfluss auf die Item-Schwierigkeit. Negative Fragen können die Frage aber auch *auf den Kopf* stellen und damit den Denkprozess um *einen Grad* erschweren. Stärker ist dieses Phänomen, wenn man dann auch noch doppelte Verneinungen in der Formulierung hat.

Typ-K Fragen sind schwerer ($p<0{,}001$ und $d=0{,}16$), da sie eine Lösungsebene mehr besitzen. Es werden in der Regel vier Aussagen angeboten, diese sind auf Korrektheit hin zu überprüfen, allerdings ist auf der zweiten Ebene auch die Kombination der korrekten Aussagen ein Thema. Das *Aussagen-basierte Fragen* schwerer sind, ist ebenfalls nicht verwunderlich ($p<0{,}001$ und $d=0{,}23$), da sie nicht fokussiert sind, dies wurde bereits beschrieben.

Es handelt sich um ein *Zusammenwürfeln* heterogener Aussagen, von denen eine korrekt und vier inkorrekt sein müssen. Unter diesem Umstand – nämlich dem Fehlen des Fokus – ist das Auffinden und Verwenden herausfordernder Aussagen leichter. Bei *Vignettenfragen* liegen wie bereits gesagt komplexe Sachverhalte vor, die aber eindeutig zu beantworten sein müssen. Um hier strittige Fragen zu vermeiden, werden die Fragen noch leichter weil eindeutiger formuliert ($p<0{,}001$ und $d=-0{,}35$).

Zum Vergleich dienen beispielsweise die Daten von Phipps: Seine Fall-basierten Fragenitems haben einen mittleren Schwierigkeitsgrad von 0,765 im Vergleich zu den nicht Fall-basierten Fragen mit 0,769. Hier ist der Unterschied sehr gering und bringt geringfügig schwierigere Fall-basierte Fragen. Im Vergleich dazu sind Fragen des Typs-K deutlich schwerer mit einer Schwierigkeit von 0,720 (Phipps & Brackbill 2009).

Parallel dazu sind Typ-A negative Fragen signifikant weniger trennscharf ($p<0{,}033$ und $d=0{,}09$). Hingegen signifikant trennschärfer sind Fragen des Typs-K ($p<0{,}001$ und $d=-0{,}25$), Aussagen-basierte Fragen ($p<0{,}001$ und $d=-0{,}16$) und Fragen mit Abbildungen ($p<0{,}001$ und $d=-0{,}49$).

Abschließend wird der Einfluss der Fragetypen und Item-Eigenschaften über Abschnitte hinweg, anhand der Effektstärken detaillierter dargestellt: Die Effektstärken der Vignettenfragen zeigen deutlich leichtere und weniger trennscharfe Items (Abbildung 26, Abbildung 27). Wenn man sich die Effektstärken der unterschiedlichen Abschnitte ansieht, zeigt sich jedoch sehr konstant, dass Fragenitems mit Vignetten leichter sind (Abbildung 28). Die Effekte sind bei negativen Fragen ebenfalls konstant, wenn auch deutlich geringer, bei Fragen des Typs-K und bei Aussagen unterschiedlicher.

Die Effektstärken der Trennschärfe zeigen in Summe wenig Konstanz, die Werte sind in Summe aber auch sehr gering (Abbildung 29). Lediglich die Typ-K Fragen der Zwischenklinik sind auffällig wenig trennscharf und stechen etwas heraus.

Wenn man sich die Verteilung dieser Typen und Eigenschaften auf die drei Abschnitte nochmals vergegenwärtigt, wird verständlich, warum die Prüfungen der ersten Hälfte des Curriculums tendenziell schwerer und die Prüfungen der zweiten Hälfte tendenziell leichter sind.

7.2.5. Einfluss der Fragetypen auf die Distraktoren

Fragen des *Typs-K* haben relativ gute Distraktoren, ähnlich den *Aussagen-basierten Fragen*. Der Umstand, dass mehr als eine Aussage richtig sein können und die zusätzlich vorhandene Kombinatorik beim Typ-K erleichtern das Schreiben von Distraktoren. Ähnlich ist es beim Schreiben der oft sehr unterschiedlichen und heterogenen Aussagen bei Aussagen-basierten Items.

Negative Fragen haben eher schlechte Distraktoren. Ein guter Distraktor muss sinnvoll klingen und inhaltlich inkorrekt sein, diese Gradwanderung ist nicht immer leicht. Bei negativen Formulierungen hat man diese Gradwanderung nicht viermal, was die Qualität verbessern könnte.

In Wirklichkeit flüchten sich viele Autorinnen und Autoren in das Erstellen negativer Fragen, um diese Gradwanderung zu umgehen – allerdings sind es eher wenig geschulte Personen, die diese Flucht absolvieren müssen.

Warum *Vignettenfragen* sehr leichte Distraktoren haben, wurde bereits oben erläutert. Vor allem die sehr eingeschränkten Möglichkeiten erschweren das Schreiben guter Distraktoren. Wenn man beispielsweise Eigenschaften von Krankheitserregern erfragt, hat man viele Möglichkeiten. Wenn man aber beispielsweise in der Klinik sich eine alternative Erklärung zur Eintrübung der Linse im Auge eines älteren Patienten überlegt, tut man sich deutlich schwerer.

Es gibt viele Elemente im Periodensystem, sehr viele Bakterien und Viren, aber nur wenige alternative Optionen zu sehr spezifischen Sachverhalten in Teilbereichen der klinischen Fächer.

Über alle Fragetypen und Item-Eigenschaften hinweg gibt es eine signifikante Beeinflussung der Distraktoren-Qualität: abgesehen von den Fragen mit Rechnungen (die keinen signifikanten Unterschied zeigen) und der negativen Typ-A Fragen, die nur einen gering signifikanten Unterschied zeigen, haben alle anderen einen hoch signifikanten Unterschied.

Passend zur hohen Anzahl von zu leichten Fragen, sind sehr viele Fragenitems – unabhängig des Fragetyps und der Eigenschaften – mit schlechten Distraktoren ausgestattet. Dieser Umstand ist allerdings nur bei Vignettenfragen und negativen Fragen besonders ausgeprägt (Abbildung 30). Sehr schlechte Distraktoren gibt es glücklicherweise seltener, aber auch hier schneiden Vignettenfragen und negative Fragen schlechter ab (Abbildung 31).

7.2.6. Schlüsselaussagen unter Einbeziehung der PTM-Ergebnisse

Wenn man diese Ergebnisse zusammen sieht und interpretiert, kann man unter anderem folgendermaßen zusammenfassen:

Die spezifischen Fragetypen und Frageeigenschaften sind in ihrem Vorkommen großteils sehr einseitig auf die drei Abschnitte verteilt. Da diese Eigenschaften oft signifikant unterschiedliche Kennwerte im Sinne von Item-Schwierigkeit und Item-Trennschärfe bedingen, haben auch die drei Abschnitte teils unterschiedliche Kennwerte. Aussagen-basierte Fragen, aber auch noch immer zu findende Fragen des Typs-K, heben den Schwierigkeitsgrad im vorklinischen Bereich deutlich an. Fragen mit Fallvignetten, aber auch negative Formulierungen, lassen die Schwierigkeit und auch die Trennschärfe in der Klinik absinken.

Ob die komplexen Inhalte der anwendungsorientierten und praxisnahen Fächer der Klinik nun direkt diese Auswirkungen auf die Schwierigkeit haben, oder die Fragetypen bedingen, die wiederum Auswirkungen auf die Schwierigkeit haben, kann nicht eindeutig beantwortet werden.

Daneben muss auch betont werden, dass alle Mittelwerte der Item-Trennschärfe und besonders der Item-Schwierigkeit verbesserungswürdig sind. Auch wenn die Klinik und Vignettenfragen schlechter abschneiden, so sind auch die Kennzahlen der anderen Abschnitte und Fragetypen nicht überwältigend. So hat beispielsweise auch die Vorklinik einen Median der Item-Schwierigkeit von über 0,80 – was bedeutet, dass die Hälfte der Fragenitems eine Schwierigkeit von über 0,80 haben muss.

Bei der Zwischenklinik liegt der Median bei 0,75 – wäre aber auch verbesserungswürdig, wenn man sich vor Augen hält, dass die optimale Trennung von guten und schlechten Studierenden bei einer Schwierigkeit von 0,50 stattfindet.

Analog dazu hat der Mittelwert der Item-Trennschärfe einen Wert von 0,24 über alle Abschnitte hinweg. Ein Wert den Möltner als *akzeptabel* ansieht, der Wert von 0,20 in der Klinik oder sogar 0,17 im zweiten Studienjahr bewertet er als *marginal* (Möltner et al. 2006)

Die Vorklinik, die – nicht im negativen Sinne – isolierte Fakten abfragt, tut sich beim Schreiben präziser und auch hinreichend schwerer Fragen leichter, als die klinischen Fächer, deren komplexe Themen nur mühsam in präzise und schwere Fragen zu packen sind. Dieser Umstand könnte ein Faktor zur Erklärung der PTM-Ergebnisse und dem schlechteren Abschneiden in den klinischen Jahren sein. Die Belastbarkeit der PTM-Ergebnisse vorausgesetzt, gibt es auch noch andere Faktoren, beispielsweise die zeitliche Positionierung der jeweiligen Inhalte im Curriculum.

Beim Vergleich der Ergebnisse der Fragenanalyse und der PTM-Ergebnisse muss man aber auch anmerken, dass die anderen Fakultäten im Vergleich dieselben Probleme haben müssten. Damit würden die PTM-Leistungen der Medizinischen Universität Graz und der anderen ähnlich und vergleichbar sein, was nicht der Fall ist. Eine Erklärung könnte sein, dass die klinischen Fragen in ihrer Formulierung und Qualität nicht die passende Begründung sind. Alternativ könnte man aber auch vermuten, dass die Medizinische Universität Graz schlechter mit diesen anwendungsorientierten Fragen umgeht. Es könnte auch sein, dass vergleichbare Fakultäten und Universitäten in Mitteleuropa (die die Vergleichsgruppe im PTM waren) andere Fragetypen einsetzen oder diese Fragenitems besser formulieren, längere Tradition mit diesen Fragetypen haben oder überhaupt andere Prüfungsformate verwenden.

Anders formuliert könnte man zusammenfassen, dass die Analyse des Fragenpools zu dem Ergebnis geführt hat, dass alle Fragen verbesserungswürdig sind, allen voran aber die der Klinik. Die Analyse der PTM-Ergebnisse hat zu dem Schluss geführt, dass es vor allem in der Klinik einen Nachholbedarf gibt – wie weit dieser über die Verbesserung der MC-Fragenitems gedeckt werden kann, kann hier nicht eindeutig beantwortet werden.

7.3. GRENZEN UND EINSCHRÄNKUNGEN

Diese Arbeit hatte einen explorativen Fokus, sie sollte den gut analysierbaren Teil der aktuell verwendeten Fragenitems der Medizinischen Universität Graz untersuchen. Einerseits sollte der Fragenpool charakterisiert werden und Eigenschaften und Kennwerte auf ihre Verteilung hin untersucht werden. Andererseits sollen Zusammenhänge untersucht werden, ebenso wie der Einfluss der drei Studienabschnitte. Am Ende steht die Frage nach einem Erklärungsmodell oder zumindest einem Erklärungsansatz für die Ergebnisse des PTM.

Eine Limitation ergibt sich aus dem Umfang und der Qualität der verfügbaren Daten. Beim Umfang muss betont werden, dass Computer-basierte Prüfungen und Prüfungen mit anderen Formaten nicht mit einbezogen wurden. 13 Module verwenden MC-Fragen in einem Papier-basierten Format und wurden untersucht, daneben gibt es 9 Module, die Computer-basiert prüfen und 6 Module, die Short-Answer-Fragen haben (die Kategorien überschneiden sich). Damit wurden nur etwa 50 % aller möglichen relevanten Fragenitems untersucht. Zudem wurden nur Fragenitems untersucht, die insofern relevant sind, als sie bei Prüfungen verwendet wurden. Es wurden keine Items analysiert, die im Fragenpool vorhanden sind, aber nicht eingesetzt wurden.

Allerdings lässt sich die Einschätzung der Qualität des gesamten Fragenpools über die Prüfungsergebnisse hochrechnen. Dass sich der durchschnittliche Schwierigkeitsgrad der Fragenitems und damit der Prüfungen sowohl in der Item-Schwierigkeit, aber auch in der Anzahl korrekter Antworten der einzelnen Studierenden – und damit in der Notenverteilung – manifestiert, sollte man hier voraussetzen.

	Papier-basierte F.	Computer-basierte F.	Andere F.	Gesamt
Vorklinik	2,98	3,10		3,04
Zwischenklinik	3,41		2,92	3,17
Klinik	2,94	2,77	2,43	2,71
Alle Abschnitte	3,11	2,94	2,68	2,97

Tabelle 17: Tabellarische Auflistung der Mittelwerte der Notenverteilungsmediane aller Prüfungen der beiden untersuchten Studienjahre („F." steht dabei für „Formate").

Wenn man sich die Prüfungsergebnisse in Hinblick auf die Notenverteilungen der beiden untersuchten Studienjahre – wie in Tabelle 17 – ansieht, ergibt sich folgendes Bild: Der Mittelwert der Mediane aller Notenverteilungen dieser beiden Jahre beträgt 2,97. Im Detail zeigt sich ein ähnliches Bild wie in der durchschnittlichen Item-Schwierigkeit der drei Abschnitte, die Vor- und Zwischenklinik haben relativ schwere Prüfungen, der klinische Abschnitt ist mit Abstand der leichteste.

Wenn man Papier-basierte und Computer-basierte Prüfungsdurchführungen vergleicht, so stellt man fest, dass in der Vorklinik Computerprüfungen schwerer sind (allerdings setzt nur ein Institut diese Form ein), in der Klinik allerdings leichter (alle sechs Module untersucht). Dieser letzte Punkt ist hier besonders relevant: Der Umstand mit leichten Prüfungen und guten Prüfungsergebnissen im klinischen Abschnitt verstärkt sich sogar noch zusätzlich, wenn man Computer-basierte Formate hinzunimmt. Damit ist anzunehmen, dass bei den Fragenitems aus den Computer-basierten Prüfungen zumindest ähnliche Charakteristika vorherrschen.

Eine weitere, oben bereits erwähnte Limitation betrifft die Auswertung der Fragenitems: In dieser Arbeit wurden nur Eigenschaften und Kennwerte erhoben und untersucht, die leicht und zuverlässig, also auch *effizient* zu gewinnen waren. Wenn man sich für jede (der circa 4500) Fragen genügend Zeit nimmt, könnte man weitere interessante Eigenschaften untersuchen und dazu qualitativ beschreiben. Hätte man genügend begutachtende Personen und könnte man die Einschätzungen aller vergleichen und mitteln, hätte man viel mehr Möglichkeiten, Aussagen über die verfügbaren Fragenitems zu treffen.

Eine kleine Anmerkung ist der Zusammenstellung der Studienabschnitte gewidmet. Die verwendete Einteilung hat einige Jahre Tradition, könnte jedoch genauso von einer anderen Einteilung des gesamten Curriculums im Verlauf abgelöst werden. Gerade im Modulsystem der Medizinischen Universität Graz, in dem viele Fachabteilungen in unterschiedlicher Weise in den Modulen zusammenarbeiten, wären auch andere Einteilungen denkbar und argumentierbar.

Zu guter Letzt muss auch wiederholt festgestellt werden, dass die vorliegende Aufarbeitung des untersuchten Fragenpools nur einen kleinen Aspekt berücksichtigen kann, wenn es um einen möglichen Einfluss auf das Ergebnis der PTM-Beteiligung durch unsere Studierenden geht. Zum einen können das PTM-Ergebnis und sein Zustandekommen nicht ungefragt übernommen werden.

Ein detaillierter Einblick, wie der Test aufgebaut und vorbereitet wird, ebenso wie er durchgeführt und ausgewertet wird, ist hier absolut notwendig. Auf der anderen Seite muss man sich auch Gedanken machen, wie man die Ergebnisse im Vergleich gegenüberstellt – hier hat Muijtjens einen Überblick gegeben (Muijtjens et al. 2008).

Dennoch ist der PTM – so strittig sein Ergebnis auch sein könnte – eine wichtige und vor allem externe Referenz, wenn es darum geht, die Leistung unserer Studierenden einzuschätzen.

Andere Faktoren, die einen deutlichen Einfluss auf die Ergebnisse und damit auf deren Interpretation haben könnten, sollen nur kurz erwähnt werden: Das könnte bei der Abstimmung der Inhalte und deren Zuordnung zu einzelnen Abschnitten des Tests beginnen und bis zur Einbettung des PTM an anderen Fakultäten und damit verbunden, die Motivation der anderen Teilnehmenden, gehen. Dennoch ist das vorliegende Ergebnis einmal ein Anfang, ein erster im Detail beschriebener Faktor, dessen Kernaussagen auch in das Gesamtbild passen und plausibel sind.

7.4. ABSCHLUSS UND AUSBLICK

Wie auf der Seite 121 im letzten Absatz bereits zusammengefasst, gibt es Möglichkeiten der Verbesserung über den gesamten Fragenpool hinweg, aber natürlich mit einem Schwerpunkt im klinischen Bereich. Die Notwendigkeit dieses Schwerpunktes zeigt sich auch in den PTM-Ergebnissen.

Der aktuell vorliegende Fragenpool hat eine teils auffällige, großteils aber erklärbare Verteilung der Fragetypen und Frageneigenschaften. Negative Formulierungen sollen im Bereich der Klinik vermieden werden, Typ-K Fragen in der Vorklinik. Aussagen-basierte Fragenitems sollten adaptiert werden und am Ende fokussierter werden. Im Bereich der Klinik sollen deutlich mehr Fallvignetten eingebaut werden, wie im gesamten Curriculum mehr Abbildungen und auch klinische Fotos eingesetzt werden sollen.

Ungünstigere Eigenschaften hinsichtlich Item-Schwierigkeit, Item-Trennschärfe und der Distraktorenqualität kommen vermehrt im Bereich des klinischen Abschnitts vor. Wenn man den Hintergrund des komplexen klinischen Alltags hinzunimmt, der schwer in gute Fragen zu verpacken ist und das Abschneiden im PTM, wird die Handlungsnotwendigkeit ersichtlich.

Wenn es um die Schwierigkeit beim Generieren von Fragen geht, die höhere kognitive Fähigkeiten voraussetzen, möchte ich abschließend Tractenberg zitieren: Er meint, dass das Erstellen dieser Fragen deshalb schwer ist, da Fachexpertinnen und Fachexperten wie Novizen denken müssten (Tractenberg et al. 2013).

Es ist notwendig, an der Behebung dieser Probleme zu arbeiten – mögliche Lösungsansätze können dabei folgende Punkte sein:

- Schulungsmaßnahmen
- Betreuung beim Fragenerstellen
- Review-Prozess
- Itemanalysen und Nachbesprechungen
- …am besten lückenlos und mit Konsequenzen?!

Alternativ kann man auch einen Wechsel zu oder ein Ergänzen mit anderen Formaten überlegen. Einige Gedanken dazu: Auch wenn die Stufen der Pyramide im Bereich von „shows how“ und „does“ aktuell oft noch schlecht repräsentiert sind, kann die Stufe „knows how“ mit schriftlichen und mündlichen Formaten beurteilt werden.

Neben der offensichtlichen Variante, das Problem zu beheben und der alternativen Variante, den gesamten Prozess umzugestalten, steht die genauere Untersuchung des Problems.

Das oft zitierte Statement „Assessment drives learning“ wird auch in dieser Arbeit genutzt, um die Abschlussworte einzuleiten. Eine zuverlässige und valide Leistungsbeurteilung ist in einer Ausbildung absolut notwendig. Dabei geht es nicht vordergründig um den summativen Charakter und eine Beurteilungen mit Konsequenzen für die Studierenden, sondern auch um solide und brauchbare Rückmeldung für die Studierenden. MC-Fragen sind dafür ein effizientes und gerade im digitalen Zeitalter vielseitig einsetzbares Mittel, um diese Beurteilung vornehmen zu können.

Dabei muss aber abschließend auch ein anderer wichtiger Punkt wiederholt werden: Nicht das Prüfungsformat spielt dabei die Hauptrolle, sondern der Inhalt – gerade deshalb ist es so entscheidend, sich mit diesem auseinander zu setzen. Wie formuliere ich Sachverhalte, Probleme, Szenarien und Regeln in naturwissenschaftlichen Grundlagenfächern – und wie formuliere ich realistische Probleme aus dem klinisch-praktischen Alltag, wenn ein Patient in meiner Ordination Platz nimmt und sagt:

> *„Ich mache mir seit Tagen solche Sorgen, dass es nichts Bösartiges ist. Was meinen Sie? “*

Literaturverzeichnis

Abdulghani, H.M. et al., 2015. Faculty development programs improve the quality of Multiple Choice Questions items' writing. *Scientific reports*, 5, p.9556.

Albanese, M.A., 1982. Multiple-Choice Items with Combinations of Correct Responses: A Further Look at the Type K Format. *Evaluation & the Health Professions*, 5(2), pp.218–228.

Albanese, M.A., 1993. Type K and Other Complex Multiple-Choice Items: An Analysis of Research and Item Properties. *Educational Measurement: Issues and Practice*, 12(1), pp.28–33.

Bauer, D. et al., 2011. Pick-N multiple choice-exams: a comparison of scoring algorithms. *Advances in health sciences education* □*: theory and* 16(2), pp.211–21.

Bauer, D., Kopp, V. & Fischer, M.R., 2007. Answer changing in multiple choice assessment change that answer when in doubt--and spread the word! *BMC medical education*, 7(28).

Berendonk, C. & Beyeler, C., 2004. Strukturiertes Feedback in der ärztlichen Weiterbildung: Mini-CEX und DOPS. *Academic Medicine*, 79(10 Suppl), pp.S70–81.

Beullens, J., van Damme, B. & Jaspaert, H., 2002. Are extended-matching multiple-choice items appropriate for a final test in medical education? *Medical Teacher*, 24(4), pp.390–395.

Beullens, J., Struyf, E. & Van Damme, B., 2005. Do extended matching multiple-choice questions measure clinical reasoning? *Medical Education*, 39(4), pp.410–7.

Böhme, K. et al., 2012. Vergleich kollegialer Einzel- mit Gruppen-Reviews allgemeinmedizinischer Multiple-Choice-Fragen. *GMS Z Med Ausbild*, 29(4).

Boshuizen, H. & Schmidt, H., 1992. On the role of biomedical knowledge in clinical reasoning by experts, intermediates and .... *Cognitive Science: A Multidisciplinary Journal*, 16(2), pp.153–184.

Boursicot, K. et al., 2011. Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Medical teacher*, 33(5), pp.370–83.

Caballero, J. et al., 2014. Difficulty and discrimination indices of multiple-choice examination items in a college of pharmacy therapeutics and

pathophysiology course sequence. *The International journal of pharmacy practice*, 22(1), pp.76–83.

Cantillon, P., Irish, B. & Sales, D., 2004. Using computers for assessment in medicine. *BMJ (Clinical research ed.)*, 329(7466), pp.606–9.

Case, S.M. & Swanson, D.B., 1998. *Constructing written test questions for the basic and clinical sciences*, National Board of Medical Examiners.

Case, S.M. & Swanson, D.B., 1993. Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine*, 5(2), pp.107–115.

Case, S.M., Swanson, D.B. & Becker, D.F., 1996. Verbosity, window dressing, and red herrings: do they make a better test item? *Academic medicine : journal of the Association of American Medical Colleges*, 71(10 Suppl), pp.S28–30.

Cassels, J.R.T. & Johnstone, A.H., 1984. The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, 61(7), p.613.

Charlin, B., Tardif, J. & Boshuizen, H.P., 2000. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Academic Medicine*, 75(2), pp.182–190.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), pp.297–334.

Davis, M.H., 2003. OSCE: the Dundee experience. *Medical Teacher*, 25(3), pp.255–261.

Downing, S.M., 2004. Reliability: on the reproducibility of assessment data. *Medical Education*, 38(9), pp.1006–1012.

Downing, S.M., 2006. Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna, eds. *Handbook of Test Development*. Mahwah NJ: Lawrence Erlbaum Associates, pp. 287–301.

Downing, S.M., 2002. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Advances in health sciences education : theory and practice*, 7(3), pp.235–241.

Downing, S.M. & Haladyna, T.M., 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), pp.327–333.

Epstein, R., 2007. Assessment in medical education. *N Engl J Med*, 356(4), pp.387–396.

Epstein, R. & Hundert, E., 2002. Defining and assessing professional competence. *JAMA*, 287(2), pp.226–235.

Fabrey, L.J. & Case, S.M., 1985. Further support for changing multiple-choice answers. *Journal of medical education*, 60(6), pp.488–490.

Fischer, M.R., Kopp, V., et al., 2005. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. *Medical Teacher*, 27(5), pp.450–455.

Fischer, M.R., Herrmann, S. & Kopp, V., 2005. Answering multiple-choice questions in high-stakes medical examinations. *Medical Education*, 39(9), pp.890–894.

Frey, P., 2006. Computerbasiert prüfen: Möglichkeiten und Grenzen Computer-based Assessment: Potentials and Drawbacks. *GMS Z Med Ausbild*, 23(3), p.Doc49.

Frisbie, D.A., 1992. The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, 11(4), pp.21–26.

Hager, P. & Gonczi, A., 1996. What is competence? *Medical Teacher*, 18(1), pp.15–18.

Haladyna, T.M. & Downing, S.M., 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), pp.309–334.

Haladyna, T.M. & Downing, S.M., 1989. A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), pp.37–50.

Haladyna, T.M. & Downing, S.M., 1989. Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2(1), pp.51–78.

Harden, R.M., 2000. Evolution or Revolution and the Future of Medical Education: Replacing the Oak Tree. *Medical Teacher*, 22(5), pp.435–442.

Hochlehnert, A. et al., 2012. Good exams made easy: the item management system for multiple examination formats. *BMC medical education*, 12(1), p.63.

Holtzman, K.Z. et al., 2009. Use of multimedia on the step 1 and step 2 clinical knowledge components of USMLE: a controlled trial of the impact on item characteristics. *Academic medicine : journal of the Association of American Medical Colleges*, 84(10 Suppl), pp.S90–3.

Huddle, T.S. & Heudebert, G.R., 2007. Taking apart the art: the risk of anatomizing clinical competence. *Academic medicine : journal of the Association of American Medical Colleges*, 82(6), pp.536–41.

Ikah, D.S.K. et al., 2015. Clinical vignettes improve performance in anatomy practical assessment. *Anatomical sciences education*, 8(3), pp.221–9.

Jones, R. et al., 2001. Changing face of medical curricula. *The Lancet*, 357(9257), pp.699–703.

Kopp, V. & Möltner, A., 2006. Key-Feature-Probleme zum Prüfen von prozeduralem Wissen: Ein Praxisleitfaden. *GMS Z Med Ausbild*, 23(3), p.Doc50.

Krathwohl, D.R., 2002. A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), pp.212–218.

Krebs, R., 2004. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen.

Krebs, R., 2008. Multiple Choice Fragen? Ja, aber richtig.

Kropf, R. et al., 2010. Auswirkungen angeleiteter Itemanalysebesprechungen mit Dozierenden auf die Qualität von Multiple Choice Prüfungen. *GMS Z Med Ausbild*, 27(3).

Kubinger, K.D., 2005. Objektive psychologisch-diagnostische Verfahren. In H. . R. T. Weber, ed. *Handbuch der Persönlichkeitspsychologie und Differentiellen Psychologie*. Göttingen: Hogrefe, pp. 158–165.

Kuder, G.F. & Richardson, M.W., 1937. The theory of the estimation of test reliability. *Psychometrika*, 2(3), pp.151–160.

Lienert, G.A. & Raatz, U., 1998. *Testaufbau und Testanalyse* 6. Auflage., BeltzPVU Weinheim.

Mann, K. V, 2011. Theoretical perspectives in medical education: past experience and future possibilities. *Medical Education*, 45(1), pp.60–68.

McLachlan, J.C., 2006. The relationship between assessment and learning. *Medical education*, 40(8), pp.716–7.

Miller, G.E., 1990. The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(9).

Millman, J., Bishop, C.H. & Ebel, R., 1965. An analysis of test-wiseness. *Educational and Psychological Measurement*, 25(3), pp.707–726.

Möltner, A., Schellberg, D. & Jünger, J., 2006. Grundlegende quantitative Analysen medizinischer Prüfungen Basic quantitative analyses of medical examinations. *GMS Z Med Ausbild*, 23(3).

Muijtjens, A.M.M. et al., 2008. Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical education*, 42(1), pp.82–8.

Newble, D., 2004. Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education*, 38(2), pp.199–203.

Nnodim, J.O., 1992. Multiple-choice testing in anatomy. *Medical Education*, 26(4), pp.301–309.

Norcini, J.J., 2003. Setting standards on educational tests. *Medical Education*, 37(5), pp.464–469.

Norcini, J.J. et al., 1995. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Annals of Internal Medicine*, 123(10), pp.795–799.

Norcini, J.J. & McKinley, D.W., 2007. Assessment methods in medical education. *Teaching and Teacher Education*, 23(3), pp.239–250.

Öchsner, W. & Böckers, A., 2013. Medical journals as implicit role models for good multiple choice questions: to which degree have CME tests so far fulfilled formal quality criteria? *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 107(7), pp.468–74.

Page, G. & Bordage, G., 1995. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Academic Medicine*, 70(2), pp.104–110.

Palmer, E. & Devitt, P., 2007. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC medical education*, 7(49).

Phipps, S.D. & Brackbill, M.L., 2009. Relationship between assessment item format and item performance characteristics. *American journal of pharmaceutical education*, 73(8), p.146.

Ripkey, D.R., Case, S.M. & Swanson, D.B., 1996. A "new" item format for assessing aspects of clinical competence. *Academic Medicine*, 71(10 Suppl), pp.S34–6.

Rodriguez, M.C., 1997. The art & science of item writing: A meta-analysis of multiple-choice item format effects. In *Annual meeting of the American Education Research Association, Chicago, IL*.

Rodriguez, M.C., 2005. Three Options Are Optimal for Multiple Choice Items: A Meta Analysis of 80 Years of Research. *Educational Measurement: Issues and*, 24(2), pp.3–13.

Rotthoff, T. & Soboll, S., 2006. Qualitätsverbesserung von MC Fragen - Ein exemplarischerWeg für eine medizinische Fakultät. *GMS Z Med Ausbild*, 23(3), p.Doc45.

Schneid, S.D. et al., 2014. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Medical education*, 48(10), pp.1020–7.

Schulze, J., Drolshagen, S. & Nürnberger, F., 2005. Einfluss des Fragenformates in Multiple-choice-Prüfungen auf die Antwortwahrscheinlichkeit: eine Untersuchung am Beispiel mikrobiologischer Fragen. *GMS Z Med Ausbild*, 22(4).

Schuwirth, L. & Van der Vleuten, C.P.M., 2004. Merging views on assessment. *Medical Education*, 38(12), pp.1208–1210.

Schuwirth, L.W.T., 1999. How to write short cases for assessing problem-solving skills. *Medical Teacher*, 21(2), pp.144–150.

Schuwirth, L.W.T. & Verheggen, M.M., 2001. Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education*, 35, pp.348–356.

Schuwirth, L.W.T. & Van der Vleuten, C.P.M., 2006. A plea for new psychometric models in educational assessment. *Medical Education*, 40(4), pp.296–300.

Schuwirth, L.W.T. & van der Vleuten, C.P.M., 2006. Challenges for educationalists. *BMJ (Clinical research ed.)*, 333(7567), pp.544–6.

Schuwirth, L.W.T. & Van der Vleuten, C.P.M., 2004. Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, 38(9), pp.974–979.

Scriven, M., 1991. *Evaluation thesaurus* 4th ed., Newbury Park, CA: Sage Publications.

Sibert, L. et al., 2005. Online clinical reasoning assessment with the Script Concordance test: a feasibility study. *BMC medical informatics and decision making*, 5(18).

Smolle, J., 2008. *Klinische MC-Fragen rasch und einfach erstellen*, Berlin, Boston: De Gruyter.

Swing, S.R., 2007. The ACGME outcome project: retrospective and prospective. *Medical teacher*, 29(7), pp.648–54.

Tamir, P., 1993. Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation*, 19(3), pp.311–325.

Tarrant, M. & Ware, J., 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2), pp.198–206.

Tarrant, M., Ware, J. & Mohammed, A.M., 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC medical education*, 9, p.40.

Tractenberg, R.E. et al., 2013. Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in health sciences education : theory and practice*, 18(5), pp.945–61.

Van der Vleuten, C.P.M., 1996. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*, 1(1), pp.41–67.

Vyas, R. & Supe, A., 2008. Multiple choice questions: a literature review on the optimal number of options. *The National medical journal of India*, 21(3), pp.130–133.

Wass, V. et al., 2001. Assessment of clinical competence. *The Lancet*, 357, pp.945–949.

Weih, M. et al., 2009. Qualitätsverbesserung von Multiple-Choice-Prüfungen. *Der Nervenarzt*, 80, pp.324–328.

Weih, M. & Abrahamson, S., 2006. Abrahamsons Erkrankungen des Curriculums und mögliche Therapien. *GMS Z Med Ausbild*, 23(2), p.Doc33.

Wilkinson, J.R. et al., 2008. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42(4), pp.364–373.

Wragg, A. et al., 2003. Assessing the performance of specialist registrars. *Clinical Medicine, Journal of the Royal College of Physicians*, 3(2), pp.131–134.