Auswertungsverfahren für Prüfungen mit multiple response-Aufgaben auf der Grundlage der Signalentdeckungstheorie

Dissertation

zur Erlangung des
Doktorgrades der Philosophie (Dr. phil.)

vorgelegt

der Philosophischen Fakultät I
Sozialwissenschaften und historische Kulturwissenschaften

der Martin-Luther-Universität
Halle-Wittenberg,

von
Herr Andreas Melzer
geb. am 10.10.1982 in Weißenfels

Erstgutachter:	Prof. Dr. Josef Lukas
Zweitgutachter:	Prof. Dr. Dieter Heyer
Tag der Verteidigung:	13. Oktober 2016

> *Religion, politics and formula scoring are areas where two informed people often hold opposing ideas with great assurance.*

Frederic. M. Lord, *1975*, S. 7

> *There is hardly a field in psychology in which the effects of signal detection theory have not been felt.*

Donald McNicol, *2005*, S. v

Danksagung

Die Anfertigung einer Dissertationsschrift ist eine zeit- und nervenaufreibende Angelegenheit. Umso wertvoller sind daher die Menschen, die den Autor dabei begleiten. Ihnen möchte ich an dieser Stelle meinen Dank aussprechen.

Zuallererst gilt mein Dank Prof. Dr. Josef Lukas für das entgegengebrachte Vertrauen und die stetige Unterstützung, die er mir bereits während meiner Studienzeit zuteil werden ließ. Allen Zweifeln und Zweiflern, die sich im Laufe der Arbeit an dieser Dissertation offenbarten, trat er mit seiner optimistischen Zuversicht entgegen und half, diese zu zerstreuen und zu entkräften.

Auch bei Prof. Dr. Dieter Heyer möchte ich mich bedanken. Er hat mich schon während des Studiums für die Signalentdeckungstheorie und R begeistert und einen wertvollen Grundstein für meine Kenntnisse in diesem Bereich gelegt. Umso erfreuter war ich, als er sich ohne Zögern als Gutachter dieser Arbeit zur Verfügung gestellt hat.

Für angeregte Diskussionen zum Thema und ausgezeichneter Hilfe beim „letzten Schliff“ an dieser Arbeit bedanke ich mich bei Sören Much, M.Sc.

Ich danke meiner Frau Juliane und meinen Kindern, die die oft mit nach Hause gebrachte Arbeit geduldig ertrugen. Auch den vielen, manchmal etwas entnervenden, Nachfragern aus der gesamten Familie, wann die Dissertation denn nun endlich fertig sei, sei gedankt. Ganz speziell seien hier meine Eltern hervorgehoben, die neben all den Dingen, die Eltern eben so machen, mit äußerster Akribie (hoffentlich) auch den letzten Rechtschreib- oder Grammatikfehler aufgespürt haben.

Nicht zuletzt gilt mein Dank Herrn Prof. Joachim Neumann, der mir freundlicherweise seine Studierenden für meine Untersuchungen zur Verfügung stellte und auch eben jenen Studierenden, die sich davon überzeugen ließen, nach einer Prüfung gleich nochmal meine Prüfungsaufgaben zu bearbeiten.

Andreas Melzer

Halle (Saale), im Januar 2016

Zusammenfassung

In der vorliegenden Arbeit soll überprüft werden, ob und inwieweit die Signalentdeckungstheorie geeignet ist, für die Auswertung von Klausuraufgaben aus *multiple response*-Prüfungen herangezogen zu werden.

Dazu wird zunächst der Begriff „*multiple response*-Aufgaben“ im Kontext von universitären Prüfungen geklärt und ein Überblick über die mögliche Ausgestaltung dieser Art von Aufgaben gegeben. Anschließend wird die Idee und das statistische Modell der Signalentdeckungstheorie vorgestellt und aufgezeigt, wie sich dieses auf *multiple response*-Aufgaben übertragen lässt.

Die Möglichkeiten der Signalentdeckungstheorie wurden in drei Prüfungen mit Medizinstudierenden in den Jahren von 2012 bis 2014 untersucht, indem nach einer regulären bestehensrelevanten Klausur mit 30 *single response*-Aufgaben im Fach Pharmakologie und Toxikologie durch die Prüflinge jeweils zehn Zusatzaufgaben aus dem gleichen Fachgebiet mit verschiedenen *multiple response*-Antwortschlüsseln bearbeitet wurden.

Es wurden sowohl klassische Summen*scores* zur Bestimmung der Bestehensraten für die regulären Prüfungen und die Zusatzaufgaben berechnet als auch Signalentdeckungsparameter für die Zusatzaufgaben geschätzt. Die Leistungen der Prüflinge

in den Zusatzaufgaben wurden mit jenen in den regulären Klausuren verglichen.

Der gewählte Untersuchungsansatz stellte sich als problematisch heraus, da die Prüflinge entgegen der Erwartungen nicht ausreichend gut auf die Prüfungen vorbereitet waren und somit die Interpretation der gewonnenen Daten erschwert ist. Die Auswirkungen dessen, die Eignung des *multiple response*-Formats für Prüfungsaufgaben und der Signalentdeckungstheorie zu deren Auswertung werden diskutiert und ein Fazit gezogen.

Inhalt

1 Einleitung

Eine kurze Darstellung über die Entwicklung von Prüfungen 1

1.1 Historisches 1

1.2 Anforderungen an ein Prüfungssystem 3

1.3 Die Herausforderung einer sinnvollen Nomenklatur 6

1.4 Probleme bei der Verwendung von *MCQ*s 8

1.5 Ziel dieser Arbeit 10

2 Signalentdeckungstheorie

Das Modell und seine Anwendung in Prüfungen 15

2.1 Die Idee hinter der Signalentdeckungstheorie 15

2.1.1 Historische Entwicklung 15

2.1.2 Experimentelle Grundlage 17

2.2 Das Modell der Signalentdeckungstheorie 20

2.2.1 *Hits* und *false alarms* 20

2.2.2 Leistung und Antworttendenz 22

2.2.3 Modellierung des Entscheidungsprozesses 24

2.2.4 Statistische Grundlagen des Signalentdeckungsmodells . . . 26

2.2.5 Wirkung der Modellparameter 29

2.3 Das *equal-variance*-Modell 30

2.3.1 Parameterschätzung 32

2.3.2 *Receiver Operating Characteristics* 35
2.4 Das *unequal-variance*-Modell 39
2.4.1 Parameterschätzung 40
2.4.2 Verlagerung des Kriteriums mittels *Rating*-Verfahren 45
2.4.3 *Receiver Operating Characteristics* 49
2.5 Die Fläche unter der *ROC*-Kurve 51
2.6 Umgang mit „extremen“ Kategorien 52
2.7 Anwendung der Signalentdeckungstheorie in anderen Bereichen . . . 54
2.7.1 Gedächtnispsychologie 54
2.7.2 Medizin 55
2.8 Übertragung des Signalentdeckungsmodells auf Prüfungen 56

3 Empirische Untersuchungen
Prüfungen in Pharmakologie in den Jahren 2012 bis 2014 **59**
3.1 Allgemeine Bemerkungen zur Methode 59
3.1.1 Klausursituation 60
3.1.2 Klausuraufgaben 61
3.1.3 Zusatzaufgaben 62
3.1.4 Bestimmung der Bestehensgrenzen abhängig von der Ratewahrscheinlichkeit 64
3.2 Prüfung in 2012 66
3.2.1 Fragestellung und Hypothesen 66
3.2.2 Methoden 67
3.2.3 Ergebnisse 70
3.2.4 Diskussion 75
3.3 Prüfung in 2013 78
3.3.1 Fragestellung 78
3.3.2 Methoden 78
3.3.3 Ergebnisse 82
3.3.4 Diskussion 88

3.4 Prüfung in 2014 . . . 90
3.4.1 Fragestellung . . . 90
3.4.2 Methoden . . . 91
3.4.3 Ergebnisse . . . 93
3.4.4 Diskussion . . . 98
3.5 Evaluation in 2013 und 2014 . . . 99
3.5.1 Zielstellung . . . 99
3.5.2 Methode . . . 100
3.5.3 Ergebnisse der Evaluation . . . 100
3.5.4 Diskussion . . . 104

4 Allgemeine Diskussion 107
4.1 Einordnung der Ergebnisse . . . 108
4.2 Das *MR*-Format und die Antwortschlüssel . . . 110
4.3 Signalentdeckungstheorie zur Auswertung von Prüfungen . . . 113
4.3.1 Kritik an der Verwendung der Signalentdeckungstheorie . . . 113
4.3.2 Praktische Probleme bei der Prüfungsauswertung mittels der Signalentdeckungstheorie . . . 117
4.3.3 Möglichkeiten zur Verbesserung des Verfahrens . . . 120
4.4 Fazit und Ausblick . . . 121

A R-Script
Schätzung der Parameter eines Signalentdeckungsmodells 123

B Verzeichnis der Zusatzaufgaben 127
B.1 Zusatzaufgaben 2012 im *SR*-Format . . . 128
B.2 Zusatzaufgaben 2012 im *MR*-Format . . . 131
B.3 Zusatzaufgaben 2013 . . . 134
B.4 Zusatzaufgaben 2014 . . . 137

C Begleittexte zu den Zusatzaufgaben 139
C.1 Begleittext 2012 . . . 139

C.2 Begleittext 2013 140
C.2.1 zusätzlich für das *MC*-Format 140
C.2.2 zusätzlich für das *MTF*-Format 141
C.2.3 zusätzlich für das *R4*-Format 141
C.3 Begleittext 2014 142

D Aussagen auf dem Evaluationsbogen **145**

Abbildungen **147**

Tabellen **149**

Literatur **151**

1

Einleitung

Eine kurze Darstellung über die Entwicklung von Prüfungen

1.1 Historisches

Es ist seit jeher unumgänglich, dass Menschen während ihres Lebens neue Fähigkeiten erwerben oder die vorhandenen verbessern müssen, um zum Allgemeinwohl, in welcher Weise auch immer, besser beitragen zu können. Unmittelbar damit verbunden ist die mehr oder weniger formalisierte Überprüfung dessen, was bzw. wie gut etwas gelernt wurde. Während solch eine Art der Überprüfung für viele handwerklich geprägte Berufe auch schon vor einigen Jahrhunderten relativ leicht durch die Anfertigung verschiedener Werkstücke und deren Qualität nachzuweisen war, ist dies für stark geistig geprägte Berufe nicht in dieser Einfachheit möglich. Oft entschied daher Stand, Herkunft oder Leumund über z.B. die Übernahme von politischen oder Ver-

waltungsämtern und nicht die tatsächlich vorhandenen Fähigkeiten (Paulsen, 1902; Wang, 2013).

Dies ist aus heutiger Sicht zwar einerseits verständlich, jedoch andererseits in unserer Gesellschaft mit der allgemeinen Rechtsgleichheit aller Menschen nicht vereinbar: Der Zugang zu Titeln, Berufen, Ämtern u.Ä. soll geregelt sein durch die Auswahl der aufgrund ihrer Fähigkeiten am besten geeigneten Bewerber. Dies geschieht heutzutage durch das Ablegen verschiedenster Prüfungen, wobei Paulsen (1902, S. 426) darunter „die systematisch ausgeführte Ermittelung des Standes der Kenntnisse und Fertigkeiten eines Prüflings durch einen Sachverständigen“ versteht. Um also allgemeingültige, mit Anderen vergleichbare Aussagen über die Fähigkeiten eines Menschen treffen zu können, ist es notwendig, dass die Umstände einer Prüfung standardisiert sind.

Die historischen Wurzeln hierzu reichen zurück bis ins China des siebenten Jahrhunderts, wo seit dem Jahr 605 die Chinesische Beamtenprüfung zur Auswahl von Regierungsbeamten abgehalten wurde und ca. 1300 Jahre lang, mit gewissen Abwandlungen, Bestand hatte (Wang, 2013). Nach einer grundlegenden Reform im 11. Jahrhundert setzte sich dort das meritokratische Prinzip vollständig durch, so dass es praktisch jedem männlichen Bewerber, unabhängig seines Standes, nur aufgrund seiner Fähigkeiten möglich war, in die höchsten Ämter aufzusteigen. Frauen blieben diese Möglichkeiten jedoch hier, wie auch in den westlichen Kulturen, noch für mehrere Jahrhunderte verwehrt.

Demgegenüber herrschte in Europa bis hinein ins 19. Jahrhundert ein ständisch gefestigtes System vor. Erst während der Zeit der Kolonialisation und insbesondere des britischen Imperialismus ab ca. 1800 erreichten die chinesischen Ideen Europa und beeinflussten maßgeblich unser heutiges westliches Prüfungssystem. So berich-

ten Bodde (1948), Têng (1943) und Wang (2013) übereinstimmend von Schilderungen englischer Diplomaten und Handelsreisender in China über das dortige Prüfungssystem, stets verbunden mit der Forderung, dieses auch in Großbritannien umzusetzen. Dies führte dazu, dass die *British East India Company* im Jahr 1806 ein *College* in London gründete, in welchem zukünftige Verwaltungsbeamte nach chinesischem Vorbild ausgebildet werden sollten (Bodde, 1948). Aufgrund der positiven Erfahrungen in Indien und durch weitere Aufforderungen aus der Öffentlichkeit wurde ein solches System in Großbritannien auch von staatlicher Seite für die eigenen Beamten im Jahr 1855 eingeführt (Kaplan & Saccuzzo, 2009). Andere europäische Länder folgten diesem Beispiel im Laufe des 19. Jahrhunderts: Deutschland um die Jahrhundertwende 1800 (Wang, 2013) und Frankreich nach einem ersten Versuch 1791 letztendlich im Jahr 1840 (Bodde, 1948). Auch in den USA wurde im Jahr 1883 ein Prüfungssystem für Staatsbeamte nach britischem Modell eingeführt (Bodde, 1948; Kaplan & Saccuzzo, 2009).

1.2 Anforderungen an ein Prüfungssystem

Verbunden mit diesen neuen Prüfungssystemen wurde es notwendig, geeignete Prüfungsmaterialien zu entwickeln. Während es in China üblich war, philosophische Aufsätze oder Gedichte zu schreiben, entwickelten sich in der westlichen Welt zahlreiche Buchstabier-, Geographie- oder Mathematiktests, welche jedoch kaum oder überhaupt nicht standardisiert waren (Mathews, 2006). Die heutige Herangehensweise beruht auf Arbeiten von Spencer (1855) und Galton (1869), die Intelligenz als eine in der Bevölkerung ungleich verteilte Variable ansahen. Sie forderten daher, dass bestimmte allgemein akzeptierte Kriterien formuliert werden müssen, um reliable Daten über die Fähigkeiten von Personen erhalten zu können.

Aus diesen Forderungen entstand die Überlegung, dass die beobachtbaren Antworten von Personen in bestimmten Tests der Schlüssel zu den verborgenen, zugrundeliegenden Denkprozessen seien. Aus diesem Rational entwickelte Alfred Binet im Jahr 1905 einen ersten standardisierten Intelligenztest (Zazzo, 1993), mit welchem es möglich war, vergleichbare Daten zur Leistungsfähigkeit von zahlreichen Personen zu erhalten. Diese Art der Überprüfung von Wissen machte sich auch E. L. Thorndike zu nutze und entwickelte um das Jahr 1900 für seine psychologischen Forschungen Prototypen heutiger Antwort-Wahl-Aufgaben (engl. *multiple choice questions*, *MCQ*, Goodenough, 1950). Hierbei handelt es sich um Aufgaben, welche aus einem sogenannten Aufgabenstamm bzw. einer Problemstellung und mehreren gebundenen Antwortoptionen, den Alternativen, bestehen (Osterlind, 1998). Dabei ist für die einzelnen Alternativen jeweils zu unterscheiden, ob es sich im Sinne der Aufgabenstellung um richtige Antworten, also Lösungen handelt oder um in diesem Sinne falsche Antworten, den sogenannten Distraktoren (Krohne & Hock, 2007; Kubinger, 2014). Die Aufgabe der Prüflinge besteht darin, diesen Sachverhalt jeweils in geeigneter Weise anzuzeigen[1].

Der Aufgabenstamm kann unterschiedlich detailliert ausgestaltet sein. Es ist möglich nur nach einem bestimmten Fakt oder der richtigen Definition für einen genannten Begriff zu fragen. Genauso kann hier nach der Richtigkeit einer ebenfalls im Aufgabenstamm enthaltenen Aussage gefragt werden oder es wird eine Situation beschrieben, z.B. in der Medizin eine Patientenanamnese mit Daten über den Patien-

[1] In der vorliegenden Arbeit sind potentiell mögliche oder tatsächlich erfolgte, beobachtbare Bewertungen von Alternativen durch Prüflinge in „Anführungszeichen“ hervorgehoben. Diese stellen dabei zunächst ein behaviorales Datum dar und sind sowohl von der tatsächlichen Richtigkeit einer Alternative (richtig oder falsch) als auch von der tatsächlichen Korrektheit des Urteils (korrekt oder inkorrekt) zu trennen, welche jeweils in normaler Schreibweise dargestellt sind.
Beispiel: Eine Alternative stellt einen im Einklang mit dem Fragestamm falschen Sachverhalt dar. Ein Prüfling beurteilt diese Alternative als „falsch“. Sein Urteil ist daher korrekt.

ten und dessen Symptomen. Der Fragestamm könnte in diesem Fall mit der Frage nach einer geeigneten Untersuchung oder Therapie abschließen. Unter dem Aufgabenstamm folgen mindestens zwei Alternativen, wobei deren maximale Anzahl einerseits vom Aufgabenstamm abhängig sein kann (z.B., wenn nur „richtig“/„falsch“ oder „ja“/„nein“ als Antworten in Frage kommen), andererseits aber theoretisch unbegrenzt groß sein kann. In der Praxis ist jedoch die Verwendung von insgesamt maximal fünf Alternativen gängig (Case & Swanson, 2002), wobei auch Formate mit mehr Alternativen Verwendung finden, um Einfluss auf die Ratewahrscheinlichkeit zu nehmen.

Frederick J. Kelly war 1914 vermutlich der Erste, der *MCQ*s zur Beurteilung von Wissen im pädagogischen Kontext nutzte (Mathews, 2006). Diese hatten den Vorteil, dass eine große Zahl von Prüflingen kostengünstig in relativ kurzer Zeit über einen großen Teil des Lernstoffs in standardisierter Weise geprüft werden konnte. Dies äußert sich in einer hohen Auswertungsobjektivität (Rost, 2004) und hoher Reliabilität (Haladyna, 1994) dieser Prüfungen. Daher erfreuten sich *MCQ*s recht bald großer Akzeptanz und Popularität, nicht zuletzt in der amerikanischen Bevölkerung, welche eine auf den Fähigkeiten ihrer Mitglieder basierende gesellschaftliche Ordnung begrüßte (Mathews, 2006).

Der vermutlich erste großflächige Einsatz von *MCQ*s fand während des Ersten Weltkrieges durch das US-Militär statt (Yoakum & Yerkes, 1920). Hier galt es, in sehr kurzer Zeit unter Einsatz von möglichst wenig Personal, die Fähigkeiten neuer Rekruten einzuschätzen und ihnen einen Einsatzzweck zuzuordnen. Aufgrund der positiven Erfahrungen, welche das US-Militär mit dieser Art Leistungsbeurteilung auch im Zweiten Weltkrieg sammelte, entschloss sich das *National Board of Medical Examiners* (*NBME*) in den 1950er Jahren, dieses Verfahren auch für die Beurteilung von Medi-

zinstudenten und zugewanderten ausländischen Medizinern zu verwenden (Case & Swanson, 2002). Mittlerweile werden weltweit bereits seit einigen Jahrzehnten *MCQ*s in Prüfungen während der Ausbildung junger Mediziner genutzt und gewinnen auch in anderen akademischen Disziplinen aufgrund immer weiter steigender Studierendenzahlen zunehmend an Bedeutung (Madaus & O'Dwyer, 1999).

1.3 Die Herausforderung einer sinnvollen Nomenklatur

Während dabei das grundlegende Konstruktionsprinzip der Aufgaben eines *MCQ*s aus Fragestamm und Alternativen gleich blieb, wurden über die Jahre hinweg verschiedene Möglichkeiten für die Beantwortung der Fragen entwickelt und entsprechend der Reihenfolge ihrer Entwicklung mit großen Buchstaben nummeriert (Case & Swanson, 2002). Diese sind in dem schon mehrfach angeführten Standardwerk von Case und Swanson (2002) dargestellt. Basierend auf diesem Werk gibt es eine Reihe weiterer Darstellungen und Abhandlungen zur fachgerechten Gestaltung von Aufgaben für *MCQ*s (z.B. Haladyna, Downing & Rodriguez, 2002; Haladyna & Rodriguez, 2013; Krebs, 2004; Krüger, 2013; Macher, 2005; Schmidts & Lischka, 2001).

Dennoch ist heute wie damals der A-Typ das meistgenutzte Format bei der Verwendung von *MCQ*s. Bei diesem Format werden eine bestimmte Anzahl von Alternativen, in der Medizin meist fünf, vorgegeben, welche für die im Fragestamm geschilderte Situation von „völlig falsch" bis „mehr oder weniger richtig" gelten können. Die abschließend gestellte Frage fordert den Prüfling auf, diejenige Alternative auszuwählen, welche die beste Antwort in der dargestellten Situation ist. Oft wird diese Art der Fragestellung als *multiple choice* (*MC*) bezeichnet (Cronbach, 1939). Diese traditionelle, vor allem im amerikanischen Sprachraum verbreitete Bezeichnung ist jedoch ungünstig gewählt. Der Begriff *„multiple"* bezieht sich dabei auf die zweifelsohne vor-

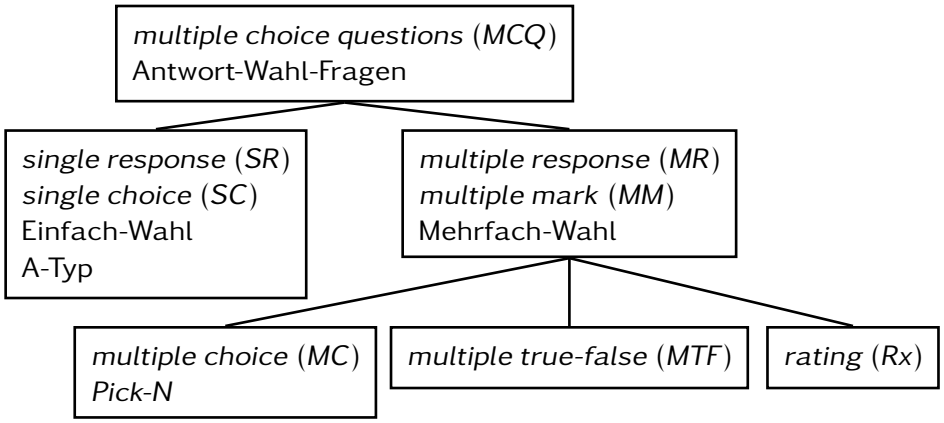

Abbildung 1.1. Schematische Darstellung der verschiedenen Möglichkeiten, Antwort-Wahl-Fragebögen (*MCQ*s) zu konzipieren (ohne Anspruch auf Vollständigkeit). Während zunächst generell zwischen Einfach-Wahl-Aufgaben (*single response, SR*) und Mehrfach-Wahl-Aufgaben (*multiple response, MR*) unterschieden werden muss, lässt sich bei Zweiterem auch das Antwortformat in unterschiedlicher Weise gestalten. Weitere Informationen hierzu finden sich im Text auf Seite 11.

handenen mehrfachen Alternativen, allerdings soll hier nur die Auswahl („*choice*“) einer einzigen Alternative getroffen werden, um korrekt zu antworten. Treffender wäre daher die Bezeichnung *single choice*- (*SC*) bzw. *single response*-Aufgabe (*SR*), während Aufgaben, bei denen mehr als eine Alternative ausgewählt werden muss, also mehrere Antworten gegeben werden müssen, als *multiple response*-Aufgaben (*MR*) bezeichnet werden sollen (s.a. Abbildung 1.1 und Frisbie, 1992).

Diese Unterscheidung zu treffen ist vor allem dahingehend wichtig, da sich Prüflinge bei der Bearbeitung von *SR*-Aufgaben sehr wohl darüber im Klaren sind, dass nur eine einzige Alternative die richtige Lösung darstellt. Es ist daher anzunehmen, dass sie diese Information strategisch nutzen, z.B. durch Anwendung eines Ausschlussverfahrens. Weiterhin begehen Prüfer implizit aber systematisch den Fehler, bei der korrekten Beantwortung einer *SR*-Frage anzunehmen, dass der Prüfling in der Lage gewesen wäre, auch die falschen Alternativen als solche zu erkennen, wenn die Frage in dieser Weise gestellt worden wäre, obwohl darüber keinerlei Information gewonnen

wird. Tatsächlich ist es sogar umgekehrt der Fall, dass dieser Antwortschlüssel keinerlei Unterscheidung dahingehend zulässt, ob ein Prüfling eine Alternative bewusst nicht angekreuzt oder aus welchen Gründen auch immer unbeabsichtigt frei gelassen hat.

1.4 Probleme bei der Verwendung von *MCQ*s

Weiterhin stellte sich die Frage, ob *MCQ*s in der Lage sind, tatsächlich das Wissen eines Prüflings zu erfassen oder ob andere grundlegendere Aspekte, wie Persönlichkeitsvariablen, einen größeren Einfluss auf das Prüfungsergebnis besitzen (Rowley, 1974). Zwei viel diskutierte Eigenschaften in diesem Zusammenhang sind die sogenannte *test wiseness*, welche die Kenntnisse eines Prüflings über die Konstruktionsprinzipien von *MCQ*s und dabei möglicherweise auftretende Fehler seitens des Prüfers oder versteckter Hinweise (Millman, Bishop & Ebel, 1965; Rogers & Yang, 1996; Runté, 2001; Towns & Robinson, 1993) in den Aufgaben beinhaltet und die *test sophistication*, welche die Menge von Erfahrungen mit bereits absolvierten *MCQ*-Prüfungen eines Prüflings bezeichnet (Erickson, 1972; McPhail, 1979; Vernon, 1938, 1962). Eng verbunden mit diesen Konzepten ist das Problem des Umgangs mit Ratemöglichkeiten bei Nicht-Wissen in *MCQ*s. Gulliksen (1950) stellt dar, dass sich Raten nicht unterbinden ließe und daher nicht verhindert werden sollte (s.a. Plumlee, 1952, 1954). Allerdings steht dem die Vermutung gegenüber, dass männliche Prüflinge, eventuell aufgrund stärker ausgeprägter Risikobereitschaft, eher dazu neigen, in *MCQ*s bei Nicht-Wissen zu raten als weibliche Prüflinge und nur aufgrund dieser Strategie und nicht aufgrund ihres Wissens bessere Ergebnisse erreichen (Ben-Shakhar & Sinai, 1991; Case, Becker & Swanson, 1993; Kubinger & Gottschall, 2007; Stanger-Hall, 2012; Zimmerman & Williams, 2003). Dies widerspräche dem Grundsatz der Gleichbehand-

lung aller Prüfling.

Daher wurden bereits in der Vergangenheit Versuche unternommen, Aufgabentypen in *MCQ*s zu verbessern (Haladyna & Downing, 1989): So wurde einerseits untersucht, wie viele Alternativen vorgegeben werden sollten (Rodriguez, 2005), ob mehrere korrekte Alternativen und ob nur korrekte oder auch falsche Alternativen verwendet werden sollten (Lienert & Raatz, 1998) und andererseits versucht, den Einfluss des Ratens herauszurechnen (Davis, 1967; Espinosa & Gardeazabal, 2010; Frary, 1969, 1988; Lyerly, 1951). Es stellte sich heraus, dass die Entwicklung von guten, sinnvollen und in etwa gleichmäßig effektiven falschen Alternativen, den sogenannten Distraktoren, oft sehr aufwendig ist (Haladyna et al., 2002) und gleichzeitig nicht garantiert ist, dass diese zur Auflösung von Fehlkonzepten beitragen. Es konnte im Gegenteil sogar gezeigt werden, dass Distraktoren, welche Prüflinge als „richtige“ Lösung auswählen, von diesen als tatsächlich richtig angenommen werden und diese Fehleinschätzung später weiterhin Bestand hat, der sogenannte *negative suggestion effect* (Brown, Schilling & Hockensmith, 1999; Preston, 1965; Roediger & Marsh, 2005; Toppino & Luipersbeck, 1993).

Dennoch sind *MCQ*s aus der heutigen Prüfungslandschaft nicht wegzudenken, da diese weiterhin den Vorteil bieten, sehr große Mengen an Prüfungsstoff bei vielen Prüflingen in kürzester Zeit zu überprüfen. Heutzutage gibt es darüber hinaus eine Vielzahl von Plattformen, welche sich dazu eignen, *MCQ*-Prüfungen elektronisch bzw. halb-elektronisch durchzuführen und auszuwerten, was zu weiterer Zeitersparnis im Prüfungsprozess führt. Ein Beispiel für die erste Variante ist ILIAS EA (ILIAS open source e-Learning e.V., 2015), ein Beispiel für die zweite Variante EvaExam (Electric Paper Evaluationssysteme GmbH, 2015a). Neben der schnellen Berechnung der erreichten Punktzahlen und ggf. sogar automatischer Benotung der einzelnen Prüflinge bietet

sich für den Prüfer hier der weitere Vorteil, dass diese Systeme in der Lage sind, quantitatives Feedback über die einzelnen Aufgaben eines *MCQ*s zu berechnen, z.B. in Form von Item-Parametern wie Schwierigkeit oder Trennschärfe. Aus diesen Werten ist es dann für den Prüfer möglich, Rückschlüsse für die eigene Lehre zu ziehen und diese an den entsprechenden Stellen zu verbessern.

1.5 Ziel dieser Arbeit

Während das bisher Dargestellte für die Erstellung von Aufgaben und zum Teil für die Durchführung von Prüfungen eine gewisse Relevanz besitzt, ist den meisten Werken gemein, dass sie dem Aspekt nach der Prüfung, nämlich der Auswertung der Aufgaben und der abschließenden Notenfindung, also der eigentlichen „systematisch(en) Ermittelung … der Kenntnisse eines Prüflings“ (Paulsen, 1902, vgl. oben), wenig bis keine Aufmerksamkeit schenken. Am häufigsten werden die Punkte der einzelnen Aufgaben mittels klassischer Summation zu einem Gesamtwert zusammengefasst und die Note aus dem Prozentwert dieser Summe an der maximal erreichbaren Punktzahl abgeleitet. Dies ist umso verwunderlicher, da die elektronische Erfassung von Prüfungsergebnissen in Kombination mit der Leistungsfähigkeit heutiger Rechnersysteme durchaus die Möglichkeit bietet, andere, aufwendigere Auswertungsmethoden anzuwenden. Beispielhaft sei hier die *Item-Response*-Theorie genannt (Embretson & Reise, 2000; Lord, 1980; Rasch, 1980; Rost, 2004), welche u.a. zur Auswertung der Ergebnisse des *Japanese-Language Proficiency Tests* (The Japan Foundation and Japan Educational Exchanges and Services, 2011, 2015) oder der bekannten PISA-Studie (Organisation for Economic Cooperation and Development, 2012, 2015) genutzt wird. Es ergeben sich daher für die vorliegende Arbeit zwei verschiedene, jedoch eng miteinander verknüpfte Ziele: Zum einen soll die Verwendbarkeit und Akzeptanz von

MR-Aufgaben in Prüfungen untersucht und zum anderen die Anwendbarkeit eines alternativen Auswertungsverfahren für derartige Prüfungen gezeigt werden.

Durch die Verwendung von *MR*-Aufgaben werden die oben genannten methodologischen Probleme der bisher beschriebenen Formate, wie falsche Schlussfolgerungen über Distraktoren und Strategieeffekte beseitigt. Hierzu werden klassische *SR*-Aufgaben entsprechend des *MR*-Ansatzes zunächst so gestaltet, dass nicht mehr nur genau eine Alternative als richtige Lösung in Betracht kommt, sondern dass bei n Alternativen k-viele Alternativen richtig sein können, wobei gilt: $0 \leq k \leq n$ und $n = 5$. Somit ist es Prüflingen nicht mehr möglich, durch z.B. ein Ausschlussverfahren die eine richtige Alternative zu erkennen oder andere Antwortstrategien (vgl. *test wiseness*) anzuwenden.

Weiterhin ist es bei dieser Art der Gestaltung der Alternativen im Gegensatz zum *SR*-Format zusätzlich möglich, den Antwortschlüssel zu variieren. Diesbezüglich wurden drei verschiedene Implementierungen für *MR*-Aufgaben (s. Abbildung 1.1) realisiert und miteinander verglichen: erstens ein dem *SC*-Format optisch sehr ähnliches *multiple choice*-Format (*MC*), bei dem tatsächlich richtige Alternativen *angekreuzt* und tatsächlich falsche Alternativen *freigelassen* werden sollen (nur dass hier im Gegensatz zum *SC*-Format ggf. mehr als ein Kreuz zu setzen ist), zweitens ein sogenanntes *multiple true-false*-Format (*MTF*), bei dem jede Alternative einzeln als entweder tatsächlich *richtig* oder tatsächlich *falsch* markiert werden soll und drittens ein Format, bei dem wiederum für jede Alternative einzeln angegeben werden soll, ob diese tatsächlich *richtig* oder tatsächlich *falsch* ist und gleichzeitig, wie sicher sich der Prüfling bei dieser Antwort ist. Es handelt sich bei letzterem also um ein *Rating*-Format, welches allgemein beliebig viele (x) Stufen haben kann (*Rx*), hier jedoch einmal mit $x = 5$ (*R5*) und einmal mit $x = 4$ Stufen (*R4*) Verwendung findet und eine

ähnliche, jedoch direktere Realisierung des Ansatzes von Kampmeyer, Matthes und Herzig (2014) darstellt.

Die drei untersuchten Antwortschlüssel unterscheiden sich also in der Anzahl ihrer unterscheidbar feststellbaren Antwortkategorien, welche ein Prüfling nutzen kann: Beim *MC*-Format lassen sich pro Alternative zwei Fälle unterscheiden, nämlich, ob ein Prüfling die Alternative „angekreuzt“ und damit für „richtig“ befunden hat oder „nicht angekreuzt“ hat. Ist die Alternative nicht angekreuzt, kann man bei diesem Format allerdings nicht erkennen, ob der Prüfling die Alternative tatsächlich für „falsch“ hält oder ob diese aus welchen Gründen auch immer nicht beantwortet wurde. Dies ist bei den beiden anderen Implementierungen nicht der Fall, da in beiden Formaten für jede einzelne Alternative eine explizite Reaktion des Prüflings auch bei „falsch“-Beurteilungen erforderlich ist. Bleiben nun Alternativen frei, so besteht im Gegensatz zum *SC*-Format oder der *MC*-Implementierung keine Unsicherheit, ob dies als „falsch“-Antwort zu deuten ist oder ob der Prüfling hier keine Antwort abgegeben hat. Dies hat den zusätzlichen Vorteil, dass Prüflinge dazu gezwungen werden, jede einzelne Alternative auf ihre Korrektheit hin zu überprüfen und entsprechend zu beurteilen, so dass deren Wissen über diese Alternativen als direkte behaviorale Daten vorliegen und nicht aus Antworten auf andere Alternativen abgeleitet werden muss.

Somit lassen sich beim *MTF*-Format drei Antwortkategorien unterscheiden, nämlich, ob die Alternative als „richtig“ oder „falsch“ beurteilt wurde oder dass sie ausgelassen wurde. Beim *Rating*-Format *Rx* hängt die Anzahl unterscheidbarer Antwortkategorien von der Anzahl der *Rating*-Kategorien *x* ab und ist stets um den Wert Eins größer, da zu *x*-vielen *Ratings* die Möglichkeit der Auslassung hinzukommt.

Als zweites Ziel dieser Arbeit soll gezeigt werden, dass zur Auswertung von Klausuren neben der klassischen Summation der Einzelpunkte die Signalentdeckungstheo-

rie (engl. *signal detection theory, SDT*, Lukas, 2006, Macmillan & Creelman, 1991, McNicol, 2005, Swets, 1996, Wickens, 2002) in mindestens gleicher Weise geeignet ist. Vorteilhaft bei einer Auswertung nach der Signalentdeckungstheorie ist, dass dieses Verfahren in der Lage ist, die Leistung einer Person unabhängig von ihrer Antwortneigung festzustellen. Auf den Prüfungskontext bezogen bedeutet dies, dass mittels Signalentdeckungstheorie das Wissen eines Prüflings gemessen werden kann, unabhängig davon, ob und in welcher Weise dieser über *test wiseness* oder *test sophistication* verfügt, da es für die Beantwortung der einzelnen Items keine explizierbare Strategie gibt. Zusätzlich ist es mit diesem Verfahren möglich, *Rating*-Formate ohne vorherige Dichotomisierung der Kategorien auf richtig oder falsch, direkt auszuwerten.

2

Signal-entdeckungstheorie

Das Modell und seine Anwendung in Prüfungen

Im vorherigen Kapitel wurde gezeigt, dass die in einer Prüfung geschätzte Lösungswahrscheinlichkeit p_k eines Prüflings, obwohl sie fast immer benutzt wird, im Allgemeinen nicht mit dessen eigentlich zu bestimmenden Wissen p_W übereinstimmt. Daher soll nun ein Verfahren vorgestellt werden, dessen Nutzen zur Auswertung von Prüfungen im Rahmen dieser Arbeit mit dem klassischen summativen Ansatz verglichen wird: die Signalentdeckungstheorie.

2.1 Die Idee hinter der Signalentdeckungstheorie

2.1.1 Historische Entwicklung

Die Signalentdeckungstheorie ist eine klassische Methode in der Psychologie, welche zur Auswertung von Daten entwickelt wurde, die aus Experimenten stammen, in

denen Versuchspersonen uneindeutige Stimuli kategorisieren sollen. Ziel war es, unabhängig von der allgemeinen Antworttendenz einer Versuchsperson, feststellen zu können, inwieweit diese in der Lage ist, zwischen Reizmustern, welche Information beinhalten (den Stimuli bzw. Signalen) und anderen, zufälligen Mustern, welche die Information überlagern oder dieser sehr ähnlich sind (dem Rauschen, engl. *noise*), zu unterscheiden.

Diese Überlegungen gehen auf ein praktisches Problem zurück, welches sich während des Zweiten Weltkriegs für das US-Militär stellte und von Marcum (1947) beschrieben wird: Die verwendeten Radargeräte zur Luftraumüberwachung waren in ihrer Reichweite, Auflösung und Genauigkeit begrenzt. Dies führte dazu, dass neben den eigentlich interessierenden Radarsignaturen von Flugzeugen beständig zufälliges Rauschen aufgezeichnet wurde. Aufgrund der zufälligen Natur dieses Rauschens war es möglich, dass es so stark anstieg, dass es mit einem echten Signal verwechselt werden konnte. Da die Radarschützen jedoch zu jeder Zeit einschätzen mussten, ob das, was sie auf dem Radarschirm sehen, ein Flugzeug (das Signal) oder etwas anderes (Rauschen) darstellt, war es ihnen nur mit einer gewissen Wahrscheinlichkeit möglich, tatsächliche Signale zu entdecken und Fehleinschätzungen waren unumgänglich. Dennoch sollte verständlicherweise die Entdeckungsleistung möglichst hoch sein, während gleichzeitig Fehler minimiert werden sollten. Es zeigte sich jedoch bald, dass dies nicht unabhängig voneinander möglich war und zusätzlich von der Reaktionsneigung der einzelnen Person abhing.

Während sich der Bericht von Marcum (1947) stark auf den militärischen Hintergrund bezieht, entwickelten Peterson, Birdsall und Fox (1954) und van Meter und Middleton (1954) dazu eine weiterführende mathematische Theorie über die „Entdeckbarkeit“ von Signalen, welche auf grundlegenden Arbeiten zur statistischen Entschei-

dungstheorie (z.B. Grenander, 1950; Neyman & Pearson, 1933; Wald, 1947, 1950) basiert (Green & Swets, 1966). Diese mathematische Theorie wurde von Tanner und Swets (1954a) und Swets, Tanner und Birdsall (1961) aufgegriffen und in eine psychologische Theorie überführt, die den zugrundeliegenden Entscheidungsprozess beschreibt und die Überlegungen aus frühen psychophysischen Arbeiten von Thurstone (1927a, 1927b) aufgriff. Nach einigen frühen Arbeiten zur Anwendbarkeit der Signalentdeckungstheorie in psychophysischen Experimenten (z.B. Marill, 1956; Munson & Karlin, 1954; Smith & Wilson, 1953; Tanner & Swets, 1953, 1954b) folgte eine Dekade, in welcher die Signalentdeckungstheorie in einer Vielzahl derartiger Experimente Anwendung fand und von denen ein großer Teil in Swets (1964) zusammenfassend veröffentlicht wurde.

2.1.2 Experimentelle Grundlage

Die grundsätzliche experimentelle Situation in dieser Art psychophysischer Experimente, wie sie Green und Swets (1966) und nachfolgend viele andere Arbeitsgruppen (z.B. Broadbent & Gregory, 1963; A. Treisman & Geffen, 1967; M. Treisman & Watts, 1966) nutzten, ist wie folgt: In jedem Durchgang[2] wird der Versuchsperson entweder ein Signal, z.B. ein sehr leiser Ton oder ein schwaches Licht, gemeinsam mit einem maskierenden Rauschen oder nur das Rauschen präsentiert. Aufgabe der Versuchsperson ist es, für jeden Durchgang anzugeben, ob das Signal vorhanden war oder nicht. Zusätzlich kann die Intensität des Signals variiert werden, um die Absolutschwelle für eine Modalität zu bestimmen (vgl. Fechner, 1860; Klein & Macmillan, 2001; Swets, 1961).

[2] Ob einzelne Durchgänge durch die Versuchsperson voneinander abgrenzbar sind oder ein konstanter Strom von Rauschen präsentiert wird, welcher an zufälligen Zeitpunkten das Signal beinhaltet und konstant überwacht werden muss, ist von Experiment zu Experiment verschieden, aber für den theoretischen Aspekt nicht weiter von Belang.

In der Signalentdeckungstheorie wird angenommen, dass jedes Mal, wenn der Versuchsperson ein bestimmter physikalischer Reiz, hier also das Rauschen oder das Rauschen zusammen mit dem Signal, präsentiert wird, dessen Reizenergie beginnend beim Wahrnehmungsapparat und fortgeführt durch die weiterführenden neuronalen Zentren in eine geeignete interne Repräsentation variabler Größe transformiert wird (Boneau & Cole, 1967). Für die Anwendbarkeit der Signalentdeckungstheorie spielt dabei die Quelle für die Variabilität keine Rolle. Es kann jedoch angenommen werden, dass diese ihren Ursprung entweder in der Umwelt hat, z.B. den verwendeten Geräten zur Präsentation des Signals, aufgrund von Aufmerksamkeitsverschiebungen oder zufälligen Fluktuationen im Wahrnehmungssystem auftritt oder bewusst als Teil des Experiments induziert wird (Swets, 1961). All dies hat Einfluss auf die tatsächliche Größe der internen Repräsentation, welche die subjektiv wahrgenommene Reizintensität bestimmt.

Die Versuchsperson ist innerhalb des experimentellen Kontextes gezwungen, nach der Präsentation eines Reizes eine Entscheidung zwischen den zwei klar definierten Antwortalternativen „Signal anwesend“ (kurz „S“) und „Signal abwesend“ (kurz „N“ von „*noise*“) zu treffen. Es handelt sich bei dieser Art von Experimenten also um ein klassisches *Yes*/*No*-Paradigma (Fechner, 1860; Macmillan & Creelman, 1991, 2010; McKenzie, Wixted, Noelle & Gyurjyan, 2001). Allerdings ist die als Basis für die Entscheidung zur Verfügung stehende Information unzureichend oder nicht perfekt bzw. wird bewusst ein Zustand der Unsicherheit hergestellt, so dass der Entscheidungsprozess von interferierender Information begleitet ist. Dies führt dazu, dass ein Auftreten von Fehlern bzw. Fehlurteilen nicht zu vermeiden ist.

Allerdings können Fehlurteile aus unterschiedlichen Gründen auftreten: Berichtet die Versuchsperson die Abwesenheit des Signals, obwohl es tatsächlich präsentiert

wurde, kann dies einerseits an der nicht ausreichenden Sensitivität des Sinnesorgans liegen um ein derart schwaches Signal wahrzunehmen. Andererseits ist es jedoch genauso möglich, dass die Versuchsperson das Signal zwar wahrgenommen hat, sich aber zu unsicher ist und daher mit einer positiven Antwort zögert. In beiden Fällen müsste man gleichermaßen annehmen, dass die zu bestimmende Wahrnehmungsschwelle höher liegt, so dass der präsentierte Reiz diese nicht überwinden konnte, aber nur bei Ersterem ist dies aufgrund der zu bestimmenden Leistungsfähigkeit des Sinnesorgans tatsächlich gerechtfertigt. Ebenso kann der entgegengesetzte Fall eintreten: Tatsächlich wurde kein Signal präsentiert, welches hätte wahrgenommen werden können, allerdings ist sich die Versuchsperson aufgrund aller möglichen Quellen für Rauschen dessen nicht sicher und könnte dennoch zu einer positiven Antwort tendieren.

Swets et al. (1961) befanden daher die traditionellen Methoden der Psychophysik für unzureichend und forderten, dass die Leistung einer Versuchsperson, hier die Wahrnehmungsschwelle, unabhängig vom Entscheidungsprozess zu bestimmen sei. Aus diesen Arbeiten resultierte das heute als Standardwerk zur psychologischen Signalentdeckungstheorie geltende Buch „Signal Detection Theory and Psychophysics“ von Green und Swets (1966), welches einen ersten systematischen Überblick über die Theorie, ihre Anwendbarkeit und ersten Ergebnisse liefert. Ziel war es, aus den experimentell gewonnenen Daten zwei voneinander unabhängige Parameter zu schätzen: Einerseits ein Maß für die Diskriminationsfähigkeit einer Versuchsperson zwischen Signal und Rauschen und andererseits eines, welches das Antwortverhalten der Versuchsperson widerspiegelt (Abdi, 2007). Heute werden diese beiden Parameter im Allgemeinen als die Sensitivität der Versuchsperson und als deren Antwortkriterium bezeichnet.

2.2 Das Modell der Signalentdeckungstheorie

2.2.1 *Hits* und *false alarms*

Bevor das statistische Modell der Signalentdeckungstheorie dargestellt werden kann, sollen zunächst zwei zentrale Begriffe geklärt werden, welche im Folgenden immer wieder eine Rolle spielen werden: *hits* (*H*) und *false alarms* (*FA*). Diese beiden Begriffe beschreiben zwei Ereignisse, die nach der Reaktion einer Versuchsperson bei der Präsentation von Signal bzw. Rauschen eintreten können. *Hit*, also Treffer, bezeichnet dabei das Ereignis, wenn die Versuchsperson mit „Signal anwesend“ („S“) antwortet und tatsächlich ein Signal präsentiert wurde (*S*). *False alarm*, also falscher Alarm, demgegenüber bezeichnet das Ereignis, wenn die Versuchsperson zwar auch mit „Signal anwesend“ antwortet, tatsächlich aber überhaupt kein Signal, sondern nur Rauschen (*noise* bzw. *N*) präsentiert wurde.

In klassischen psychophysischen Experimenten dienten *hits* als einziges behaviorales Datum, um die Leistung einer Versuchsperson zu bestimmen. Allerdings wird im folgenden Beispiel, welches aus Wickens (2002, S. 6ff) entnommen ist, schnell klar, warum die Betrachtung der *hits* allein nicht weder ausreichend noch zielführend ist und daher zur Entwicklung der Signalentdeckungstheorie in diesem Bereich geführt hat:

Man stelle sich ein typisches Wahrnehmungsexperiment vor, in dem es Aufgabe der Versuchsperson ist, einen schwachen Ton vor einem verrauschten Hintergrund zu entdecken (z.B. Egan, Schulman & Greenberg, 1959). In einer Sitzung wird in jeweils 100 Durchgängen nur das Rauschen präsentiert und in 100 anderen Durchgängen Signal und Rauschen gemeinsam. Die Darbietung geschieht in zufälliger Reihenfolge. In der ersten Sitzung wird der Versuchsperson mitgeteilt, dass es sehr wichtig ist,

möglichst viele Signale zu entdecken und um dazu einen Anreiz zu bieten, werden für jedes korrekt identifizierte Signal, also jeden *hit*, zehn Cent als Belohnung ausgezahlt. Später wird der Versuchsperson in einer zweiten Sitzung desselben Experiments mit dem selben Signalreiz mitgeteilt, dass es nun entscheidend ist, möglichst nur dann ein Signal zu berichten, wenn auch tatsächlich ein Signal präsentiert wurde, also *false alarms* zu vermeiden. Um nun sicherzustellen, dass die Versuchsperson dieser veränderten Instruktion folgt, wird die Auszahlung von zehn Cent für jeden *hit* gestrichen und stattdessen jeder korrekt erkannte Durchgang ohne Signal mit dieser Summe belohnt.

Als beispielhaftes Ergebnis führt Wickens (2002) an, dass in der ersten Sitzung 82 der 100 Signal-Durchgänge korrekt identifiziert wurden, es also in der Nomenklatur der Signalentdeckungstheorie 82 *hits* gab, während in der zweiten Sitzung nur noch 55 *hits* auftraten. Dies überrascht zunächst, da der Signalreiz in beiden Sitzungen derselbe war und daher das Ergebnis nicht von der Entdeckbarkeit des Signals abhängen kann. Die Begründung für das Ergebnis ist jedoch recht einfach: Es handelt sich um ein sehr schwaches Signal, welches vor einem verrauschten Hintergrund dargeboten wird. Daher kann es leicht passieren, dass Signal und Rauschen sich sehr ähnlich sind und miteinander verwechselt werden, so dass die Versuchsperson unsicher ist, was die richtige Antwort ist. Um nun die Chancen auf eine möglichst große Auszahlung zu verbessern, ist es in der ersten Sitzung von Vorteil, bei Unsicherheit oft mit „Signal anwesend“ zu antworten, da bei jeder korrekten Identifizierung die Belohnung ausgezahlt wird, bei einer falschen Antwort jedoch kein Schaden eintritt. Demgegenüber ist die Maximierungsstrategie in der zweiten Sitzung eine ganz andere, nämlich im Zweifel mit „Signal abwesend“ zu antworten, da nur korrekt identifizierte Durchgänge ohne Rauschen belohnt werden.

Tabelle 2.1. Übersicht über die vier möglichen Ereignisse *hits* (*H*), *misses* (*M*), *false alarms* (*FA*) und *correct rejections* (*CR*) in einem Experiment zur Signalentdeckungstheorie.

	Antwort	
Präsentation	„Signal anwesend" („*S*")	„Signal abwesend" („*N*")
Signal (*S*)	*HIT* (*H*)	*MISS* (*M*)
Rauschen (*N*)	*FALSE ALARM* (*FA*)	*CORRECT REJECTION* (*CR*)

2.2.2 Leistung und Antworttendenz

Um tatsächlich die Sensitivität einer Versuchsperson feststellen zu können, ist es aber wünschenswert, wenn die Antwortstrategie keinen Einfluss auf den Leistungsparameter hat. Dies ist bei alleiniger Betrachtung der *hits*, also einer Beschränkung der Auswertung auf die Signal-Durchgänge, nicht möglich. Daher ist es notwendig, auch die Antworten in den Durchgängen, in denen nur das Rauschen präsentiert wird, den sogenannten *noise*-Durchgängen, in die Bewertung der Leistung mit einzubeziehen. Im Beispiel von Wickens (2002) traten in der ersten Sitzung neben den 82 *hits* auch 46 *false alarms* auf, während in der zweiten Sitzung zwar nur noch 55 *hits* auftraten, aber auch die Anzahl der *false alarms* deutlich auf 19 zurückging.

Neben *hits* und *false alarms* können offensichtlich noch zwei weitere Ereignisse eintreten, nämlich, wenn die Versuchsperson mit „Signal abwesend" („*N*") antwortet und entweder tatsächlich nur Rauschen präsentiert wurde (*N*) oder doch ein Signal (*S*). Bei Ersterem handelt es sich um eine sogenannte *correction rejection* (*CR*), bei Zweiterem um einen sogenannten *miss* (*M*). Diese vier Ereignisse lassen sich übersichtlich in einem Vierfelder-Schema wie in Tabelle 2.1 darstellen. Dabei handelt es sich bei *hits* und *correct rejections* um korrekte Antworten und bei *misses* und *false alarms* um inkorrekte Antworten. Die vollständigen Daten für beide Sitzungen aus dem Beispielexperiment von Wickens (2002) finden sich in Tabelle 2.2.

Tabelle 2.2. Vollständige Daten aus dem Beispielexperiment nach Wickens (2002, S. 8). Dargestellt sind absolute Häufigkeiten H der vier Ereignisse *hits, misses, false alarms* und *correct rejections* (S steht für Signal und N für *noise*; Anführungsstriche kennzeichen Antworten durch Versuchspersonen, ansonsten sind Präsentationen bzw. Durchgänge gemeint).

	Sitzung 1		Sitzung 2	
	„S“	„N“	„S“	„N“
S	82	18	55	45
N	46	54	19	81

In einem Experiment ist bekannt, wie viele Durchgänge mit und ohne Signal präsentiert wurden und es lässt sich feststellen, in welchen Durchgängen die Versuchsperson welche Antwort gegeben hat. Daher kann aus diesen absoluten Häufigkeiten H für die vier Ereignisse jeweils deren bedingte relative Häufigkeiten h berechnet werden:

$$\begin{aligned} h(\text{„}S\text{“} \mid S) = h_H &= \frac{H_H}{H_S} \\ h(\text{„}N\text{“} \mid S) = h_M &= \frac{H_M}{H_S} \\ h(\text{„}S\text{“} \mid N) = h_{FA} &= \frac{H_{FA}}{H_N} \\ h(\text{„}N\text{“} \mid N) = h_{CR} &= \frac{H_{CR}}{H_N} \end{aligned} \tag{2.1}$$

Da es sich bei *hits* und *misses* bzw. *false alarms* und *correct rejections* jeweils um komplementäre Ereignisse handelt, kann man sich nun leicht klarmachen, dass die folgenden Beziehungen gelten:

$$\begin{aligned} h_H + h_M &= 1 \\ h_{FA} + h_{CR} &= 1 \end{aligned} \tag{2.2}$$

weshalb es ausreichend ist, jeweils nur h_H und h_{FA} zu betrachten[3], welche auch als *hit*- bzw. *false alarm*-Rate bezeichnet werden.

2.2.3 Modellierung des Entscheidungsprozesses

Obwohl durch die gemeinsame Betrachtung von *hit*- und *false alarm*-Rate bereits viel gewonnen ist, geben diese beiden bedingten relativen Häufigkeiten allein noch wenig Auskunft über die Sensitivität einer Versuchsperson. Es ist daher wünschenswert, einen einzigen Parameter für diese bestimmen zu können. Dazu soll hier das statistische Modell der Signalentdeckungstheorie dargestellt werden, welches den Entscheidungsprozess nach der Präsentation der Reize berücksichtigt. Als Ergebnis erhält man eine Schätzung für die Sensitivität der Versuchsperson und eine zweite, davon unabhängige Schätzung für deren Antwortneigung, also ein Maß dafür, ob sie eher zu „Signal anwesend"- oder „Signal abwesend"-Antworten tendiert (Wickens, 2002).

Es ist also lohnenswert, sich den Prozess der Entscheidung in einem Durchgang nochmals zu vergegenwärtigen: Durch das Wahrnehmungssystem der Versuchsperson wird ein Reiz aufgenommen, welcher entweder nur aus zufälligem Rauschen oder aus Rauschen und dem Signal besteht. Dieser Reiz wird vom Wahrnehmungssystem über neuronale Kanäle in das Gehirn weitergeleitet und dabei ggf. transformiert. In allen Schritten dieser Verarbeitungskette können weitere zufällige Quellen für das Rauschen hinzukommen, so dass die verarbeitete Information dem Signal mehr oder weniger stark ähnelt. Diese interne Repräsentation bzw., weniger kryptisch, der subjektive Eindruck der Versuchsperson vom präsentierten Reiz, stellt die Grundlage für

[3] Prinzipiell wäre es möglich, auch ein anderes Paar bedingter relativer Häufigkeiten zu betrachten, deren beider Ereignisse sich nicht komplementär zueinander verhalten. Allerdings hat sich die Verwendung des Paares *hit*- und *false alarm*-Rate in der Literatur als Konvention durchgesetzt (Wickens, 2002).

deren Entscheidung dar, welche Art von Reiz tatsächlich präsentiert wurde. Statistisch gesehen handelt es sich hierbei um die Realisierung einer Zufallsvariablen X.

Das Modell der Signalentdeckungstheorie basiert nunmehr auf der statistischen Entscheidungstheorie (Pratt, Raiffa & Schlaifer, 1995; Wald, 1939) und macht drei Annahmen (Wickens, 2002):

1. Die Signal-Information, die die Versuchsperson einem präsentierten Reiz entnimmt, lässt sich durch eine Zahl repräsentieren.
2. Die Signal-Information ist zufälligen Schwankungen unterlegen.
3. Die Entscheidung der Versuchsperson basiert auf einer Entscheidungsregel, die die Größe der Signal-Information zugrunde legt.

Es lässt sich nun leicht erkennen, wie die Annahmen der statistischen Entscheidungstheorie auf die Annahmen über den Entscheidungsprozess im Experiment übertragen werden können: Annahme 1 über die Repräsentation der Signal-Information als einer Zahl entspricht die Realisierung der Zufallsvariablen für die interne Repräsentation. Annahme 2 über die Zufallsschwankungen der Signal-Information entspricht der zufälligen Natur des Rauschens im Verarbeitungsprozess. Annahme 3 über die Entscheidungsregel entspricht dem Verhalten der Versuchsperson, ab einer bestimmten Größe der internen Repräsentation mit „Signal anwesend“ zu antworten und ansonsten mit „Signal abwesend“.

Man kann sich daher eine kontinuierliche Dimension innerhalb der Versuchsperson vorstellen, auf welcher sich jeweils nach der Reizpräsentation der subjektive Eindruck des Reizes lokalisieren lässt, der sich wie eine Zufallsvariable X auf dieser kontinuierlichen Dimension mit einem bestimmten Erwartungswert und einer bestimmten Varianz verhält. Dabei ist dieser subjektive Eindruck umso größer ausgeprägt, je eher er dem Signal entspricht, unabhängig davon, ob tatsächlich das Signal oder diesem

sehr ähnliches Rauschen zu dieser Entsprechung führt.

Entscheidend für die Antwort der Versuchsperson in jedem Durchgang ist jeweils die Größe des subjektiven Eindrucks X. Dabei sprechen große Werte von X eher für das Vorliegen des Signals, während kleine Werte eher dagegen sprechen. Zusätzlich wird angenommen, dass die Versuchsperson intern ein festes und konsistentes Kriterium λ auf dieser Dimension generiert, welches sie für diese Entscheidung heranzieht: Solange der subjektive Eindruck X das Kriterium λ unterschreitet ($X < \lambda$) wird mit „Signal abwesend“ geantwortet und sobald dieser Eindruck X dem Kriterium λ entspricht oder größer ist ($X \geq \lambda$) mit „Signal anwesend“ geantwortet.

Dabei kann die Variation der Zufallsvariablen X im Signalentdeckungsmodell prinzipiell beliebigen Verteilungen folgen, allerdings wird oft angenommen, dass die Zufallsvariable X normalverteilt ist (Wickens, 2002). Dieser Standardfall soll auch Basis der vorliegenden Arbeit sein, wobei die statistischen Grundlagen der Normalverteilung als bekannt vorausgesetzt werden (ein Überblick findet sich bei Hays, 1994 bzw. Maxwell & Delaney, 2004).

2.2.4 Statistische Grundlagen des Signalentdeckungsmodells

Da die zur Verfügung stehende Reizinformation in Durchgängen, in denen nur das Rauschen präsentiert wird (kurz: *noise*-Durchgänge) und Durchgängen, in denen sowohl das Rauschen als auch das Signal präsentiert wird (kurz: Signal-Durchgänge) grundlegend verschieden ist, kann davon ausgegangen werden, dass die aus den jeweiligen Durchgängen extrahierte, verarbeitete Information gleichermaßen verschieden ist. Es ist daher sinnvoll, von einer Verteilung der Zufallsvariablen unter *noise* $X_n \sim N(\mu_n, \sigma_n^2)$ und einer Verteilung der Zufallsvariablen unter Signal $X_s \sim N(\mu_s, \sigma_s^2)$ zu sprechen, welche zwar auf der gleichen Dimension des subjektiven Eindrucks der

Reizinformation verteilt, jedoch o.B.d.A. an verschiedenen Orten und mit verschiedener Varianz lokalisiert sind. Dabei lässt sich das Modell jeweils so formulieren, dass größere Werte für die Anwesenheit des Signals sprechen, also $\mu_s > \mu_n$ gilt.

Es ist jedoch unmöglich, die Erwartungswerte und Varianzen der beiden Zufallsvariablen auf der Dimension des subjektiven Eindrucks genau zu bestimmen, da dieser nicht beobachtbar ist. Für die Verwendung des Signalentdeckungsmodells ist dies glücklicherweise nicht nötig, da einzig die relative Lage und Form der Verteilungen der beiden Zufallsvariablen zueinander interessiert (Wickens, 2002). Daher ist es Konsens, die *noise*-Verteilung so zu normieren, dass diese gerade der Standardnormalverteilung entspricht, also $X_n \sim N(0,1)$ gilt. Die Signal-Verteilung wird entsprechend gleichermaßen in ihrer Lage und Form angepasst, so dass sich an den relativen Verhältnissen der beiden Verteilungen zueinander nichts ändert. Dabei sollen μ_s und σ_s^2 jeweils die Parameter der Signal-Verteilung nach der Anpassung bezeichnen, also gilt: $X_s \sim N(\mu_s, \sigma_s^2)$. Diese beiden Parameter der Signal-Verteilung bilden zusammen mit dem Kriterium λ die drei ein Signalentdeckungsmodell definierenden Modellparameter. Die beiden Verteilungen von X_n und X_s eines solchen Modells sowie die Lage des Kriteriums λ auf der Dimension des subjektiven Eindrucks X sind in Abbildung 2.1 dargestellt.

Nachdem die Verteilungen für Signal und Rauschen nun festgelegt sind, lässt sich für jedes der vier möglichen Ereignisse *hits*, *false alarms*, *misses* und *correct rejections* leicht dessen Auftretenswahrscheinlichkeit bestimmen. Dazu seien hier zwei weitere Begriffe eingeführt: $\varphi(x)$ bezeichne die Dichtefunktion der Standardnormalverteilung und $\Phi(x)$ deren kumulative Verteilungsfunktion, welche beide symmetrisch sind. Damit lassen sich die Wahrscheinlichkeiten für das Auftreten von *hits* p_H bzw. *false alarms* p_{FA} als die Fläche unter der Dichtefunktion der jeweils zugehörigen Ver-

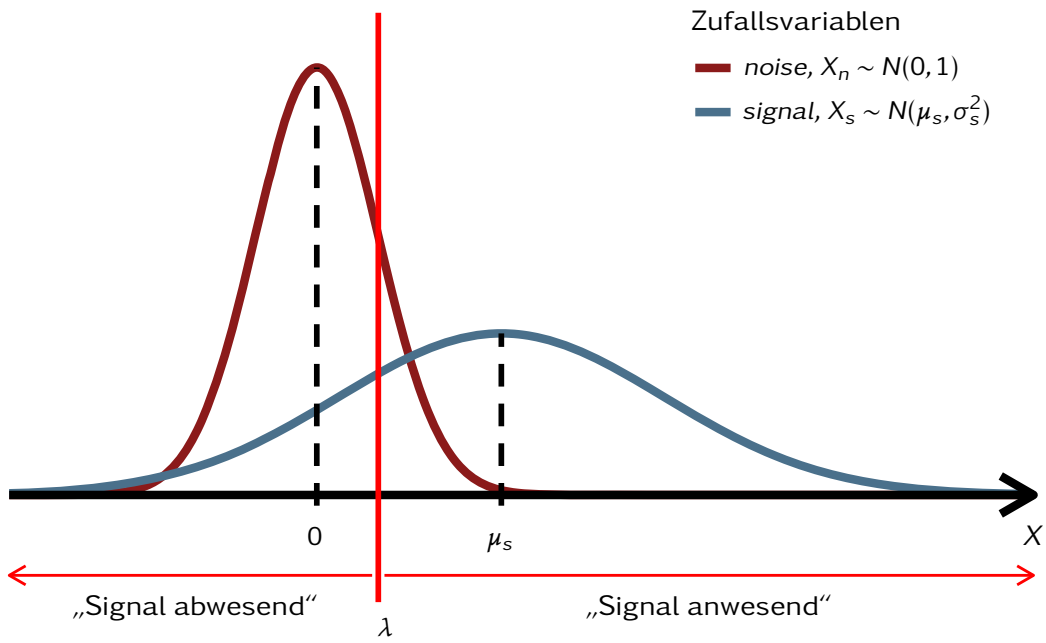

Abbildung 2.1. Dargestellt ist die Dimension des subjektiven Reizeindrucks X, die Dichtefunktionen der beiden auf dieser Dimension verteilten Zufallsvariablen X_n und X_s, das Antwortkriterium der Versuchsperson λ sowie deren Antwort je nach Größe des subjektiven Eindrucks eines Reizes.

teilung rechts vom Kriterium λ bestimmen (vgl. Abbildung 2.2). Dies lässt sich mathematisch als das Integral der Dichtefunktion $\varphi(x)$ beschreiben, wobei dieses gerade der kumulativen Verteilungsfunktion $\Phi(x)$ entspricht. Es ergeben sich:

$$\begin{aligned} p_{FA} &= P(\text{„}S\text{“} \mid N) = P(X > \lambda \mid N) = P(X_n > \lambda) \\ &= 1 - P(X_n \leq \lambda) = 1 - \int_{-\infty}^{\lambda} \varphi(x)dx \\ &= 1 - \Phi(\lambda) \end{aligned} \tag{2.3}$$

$$\begin{aligned} p_H &= P(\text{„}S\text{“} \mid S) = P(X > \lambda \mid S) = P(X_s > \lambda) \\ &= 1 - P(X_s \leq \lambda) = 1 - \int_{-\infty}^{\lambda} \varphi\left(\frac{x-\mu_s}{\sigma_s}\right)dx \\ &= 1 - \Phi\left(\frac{\lambda-\mu_s}{\sigma_s}\right) \end{aligned} \tag{2.4}$$

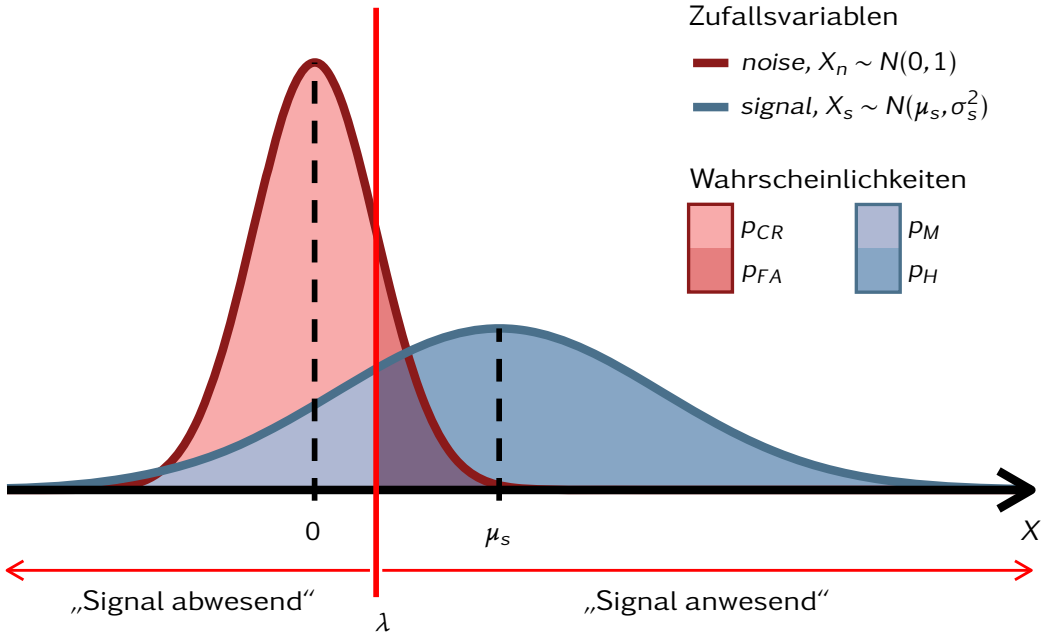

Abbildung 2.2. Dargestellt ist wiederum die Dimension des subjektiven Reizeindrucks X, die Dichtefunktionen der beiden auf dieser Dimension verteilten Zufallsvariablen X_n und X_s, das Antwortkriterium der Versuchsperson λ sowie deren Antwort je nach Größe des subjektiven Eindrucks eines Reizes.
Es ergeben sich daraus die vier möglichen Ereignisse *hit*, *miss*, *false alarm* und *correct rejection*, deren Auftretenswahrscheinlichkeiten als Flächen unter den Dichtefunktion der jeweils zugehörigen Verteilung abgetragen sind.

2.2.5 Wirkung der Modellparameter

Die statistischen Zusammenhänge zeigen, dass die *hit*- und *false alarm*-Rate unabhängig voneinander variieren können, da sie jeweils aus unterschiedlichen Grundgesamtheiten stammen. Insbesondere ist erwähnenswert, dass das Verhältnis von *hit*- und *false alarm*-Rate von zwei verschiedenen Parametern bestimmt wird, nämlich sowohl durch die Lage des Kriteriums λ als auch durch das Ausmaß der Überlappung der beiden Verteilungen, also letztlich, da die *noise*-Verteilung auf eine Standardnormalverteilung festgelegt ist, der Lage und Form der Signal-Verteilung, welche durch μ_s und σ_s^2 charakterisiert wird. Eine Veränderung des Kriteriums λ wirkt sich dabei in

gleicher Richtung, jedoch nicht im gleichen Ausmaß auf die beiden Wahrscheinlichkeiten aus: Wandert λ weiter nach links, werden sowohl *hit*- als auch *false alarm*-Rate größer, während bei einem weiter rechts liegenden Kriterium beide Wahrscheinlichkeiten kleiner werden. Man spricht in diesem Zusammenhang auch von liberaleren bzw. konservativeren Kriterien.

Demgegenüber hat eine Veränderung der Überlappung der beiden Verteilungen nur Auswirkungen auf die *hit*-Rate, da die Lage der *noise*-Verteilung und somit der Größe der *false alarm*-Rate festgelegt ist. Wandert die Signal-Verteilung bei festem Kriterium λ weiter nach rechts (μ_s wird größer) oder verringert sich deren Varianz (σ_s^2 wird kleiner), so steigt die *hit*-Rate an. Dies gibt einen Hinweis darauf, dass diese beiden Parameter dazu herangezogen werden können, um ein Maß der Entscheidungsgüte, also der Sensitivität, zu konstruieren, während die Bestimmung des Antwortkriteriums davon unabhängig ist.

2.3 Das *equal-variance*-Modell

Aus einem einfachen Signalentdeckungsexperiment erhält man als empirische Daten typischerweise nur eine *hit*- und eine *false alarm*-Rate für jede Versuchsperson. Mit diesen beiden Werten ist es jedoch nicht möglich, die drei Parameter μ_s, σ_s^2 und λ für das Signalentdeckungsmodell eindeutig zu schätzen. Es ist daher notwendig, entweder mehr Daten zu erheben oder die Menge der zu schätzenden Parameter, unter Einschränkung der Allgemeingültigkeit des Modells, weiter zu begrenzen, indem einem der Parameter ein fester, nicht empirisch belegbarer Wert zugewiesen wird. Hier soll zunächst der letztere Ansatz beschrieben werden, bevor weiter unten auf den ersteren Ansatz eingegangen wird.

Üblicherweise wählt man als jenen Parameter, dem ein fester Wert zugewiesen

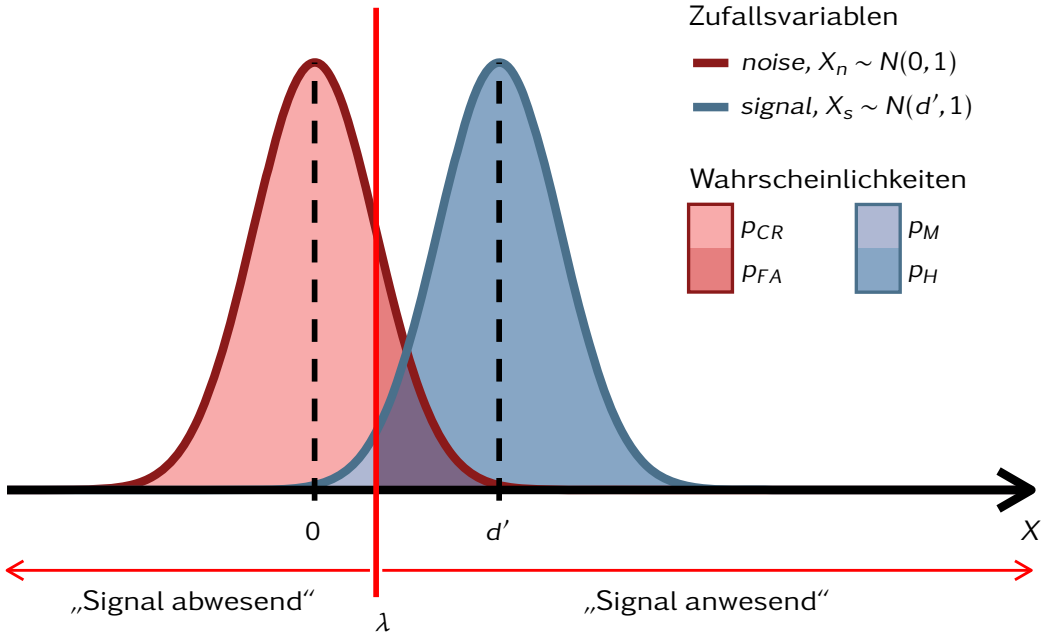

Abbildung 2.3. Dargestellt sind die bereits bekannten Dichtefunktionen der beiden auf der subjektiven Wahrnehmungsdimension normalverteilten Zufallsvariablen X_n und X_s unter der *equal-variance*-Annahme, sowie die Wahrscheinlichkeiten für die vier möglichen Ereignisse *hit*, *miss*, *false alarm* und *correct rejection* als Flächen unter den jeweils zugehörigen Dichtefunktionen.

wird, die Varianz der Signal-Verteilung σ_s^2 aus, da sich diese am schlechtesten inhaltlich interpretieren lässt, während den beiden anderen Parametern offensichtliche Funktionen zugeschrieben werden können (Wickens, 2002). Bei dieser Verfahrensweise ist es konventionell üblich, die gleiche Varianz sowohl für die *noise*- als auch die Signal-Verteilung anzunehmen, also $\sigma_s^2 = \sigma_n^2 = 1$ zu setzen, wobei aus theoretischen Gründen auch andere Werte für σ_s^2 denkbar wären. Diese genannte Variante ist in der Literatur als *equal-variance*-Modell bekannt (Macmillan & Creelman, 1991, 2010; McNicol, 2005; Wickens, 2002) und stellt einen Spezialfall des allgemeinen Modells dar. Um die spezielle Form des Modells zu verdeutlichen, wird hier der Erwartungswert der Signal-Verteilung μ_s oft mit der Bezeichnung d' versehen. Die beiden Verteilungen des *equal-variance*-Modells mit der neuen Notation sind in Abbildung 2.3 dargestellt.

Bereits in Abschnitt 2.2.5 wurde angeführt, dass sich aus der Lage μ_s und der Form σ_s^2 der Signal-Verteilung ein Maß für die Sensitivität einer Versuchsperson konstruieren lässt. Da im *equal-variance*-Modell nun nur noch der Parameter $\mu_s = d'$ zur Verfügung steht, kann dieser als ebenjenes direkte Maß für die Sensitivität einer Versuchsperson betrachtet werden.

Es ergeben sich für die Zufallsvariablen X_n für *noise* und X_s für Signal folgende Verteilungen:

$$\begin{aligned} X_n &\sim N(0,1) \\ X_s &\sim N(d',1) \end{aligned} \tag{2.5}$$

und aus den Gleichungen 2.3 und 2.4 folgt:

$$p_{FA} = 1 - \Phi(\lambda) \tag{2.6}$$

$$p_H = 1 - \Phi(\lambda - d') \tag{2.7}$$

Diese beiden Gleichungen können nun zur Bestimmung der Parameter d' und λ entsprechend umgestellt werden.

2.3.1 Parameterschätzung

Als empirische Daten werden in Signalentdeckungsexperimenten allerdings bedingte relative Häufigkeiten für die Ereignisse *hit* und *false alarm* gewonnen, aus welchen auf die beiden Parameter des zugrundeliegenden Signalentdeckungsmodells geschlossen werden muss. Es ist daher notwendig, eine Beziehung zwischen diesen bedingten relativen Häufigkeiten und den theoretischen *hit*- und *false alarm*-Raten herzustellen, bei denen es sich um die Auftretenswahrscheinlichkeiten der beiden Er-

eignisse handelt. Dies geschieht unter Anwendung des frequentistischen Wahrscheinlichkeitsbegriffs, welcher die Wahrscheinlichkeit eines Ereignisses als die relative Häufigkeit interpretiert, mit der es in einer großen Anzahl gleicher, wiederholter, voneinander unabhängiger Zufallsexperimente auftritt (Fisher, 1925; Friedman, 1999; Kendall, 1949; von Mises, 1928). Dabei spielt das Gesetz der großen Zahl (Poisson, 1837; Seneta, 2013; Tsirelson, 2012) eine entscheidende Rolle, aus welchem abgeleitet werden kann, dass sich die relative Häufigkeit eines Ereignisses ihrer theoretisch zugrundeliegenden Wahrscheinlichkeit bei vielen Wiederholungen annähert und somit die Wahrscheinlichkeit den Grenzwert der relativen Häufigkeit bei unendlich vielen Wiederholungen bildet. Dies gilt gleichermaßen für bedingte relative Häufigkeiten bzw. deren zugehörigen bedingten Wahrscheinlichkeiten.

Aufbauend auf diesen Überlegungen ist es möglich, die empirisch gewonnenen bedingten relativen Häufigkeiten für *hit* und *false alarm* als Schätzer für die ihnen entsprechenden theoretischen Auftretenswahrscheinlichkeiten dieser Ereignisse zu nutzen (Neyman, 1937). Auf dieser Grundlage können die Parameter für das Signalentdeckungsmodell geschätzt werden, indem die Gleichungen 2.6 und 2.7 entsprechend umgestellt und die theoretischen Wahrscheinlichkeiten durch ihre empirisch gewonnenen Schätzer ersetzt werden.

Um die weiteren Darstellungen sprachlich zu vereinfachen, seien von hier an die empirischen bedingten relativen Häufigkeiten für *hit* und *false alarm* wie deren zugehörige theoretische Wahrscheinlichkeiten als *hit*- bzw. *false alarm*-Rate bezeichnet, ohne dass dies gesondert gekennzeichnet wird. Dabei wird auf den Leser vertraut, den jeweils korrekten Kontext zu erkennen. Um jedoch kenntlich zu machen, dass es sich bei den auf die oben beschriebene Weise ermittelten Parametern des Signalentdeckungsmodells um empirisch geschätzte Werte handelt, werden diese mit einem

Circumflex bzw. „Dach“ versehen, also mit $\hat{\lambda}$ und $\hat{d}'$ bezeichnet.

Für das Umstellen ist weiterhin die Kenntnis der gleichfalls symmetrischen Umkehrfunktion der kumulativen Verteilungsfunktion der Standardnormalverteilung $\Phi(z)$ notwendig, wobei $z(p) = \Phi^{-1}(p)$ gilt und den Wert z für eine gegebene Wahrscheinlichkeit p auf jener Dimension liefert, auf der die zugehörige Zufallsvariable verteilt ist, im Signalentdeckungsmodell also des subjektiven Reizeindrucks. Es ergeben sich:

$$\begin{aligned} h_{FA} &= 1 - \Phi\left(\hat{\lambda}\right) \\ &= \Phi\left(-\hat{\lambda}\right) \\ z(h_{FA}) &= -\hat{\lambda} \\ -z(h_{FA}) &= \hat{\lambda} \end{aligned} \tag{2.8}$$

und

$$\begin{aligned} h_H &= 1 - \Phi\left(\hat{\lambda} - \hat{d}'\right) \\ &= \Phi\left(-\hat{\lambda} + \hat{d}'\right) = \Phi\left(\hat{d}' - \hat{\lambda}\right) \\ z(h_H) &= \hat{d}' - \hat{\lambda} \\ &= \hat{d}' - \Big(-z(h_{FA})\Big) = \hat{d}' + z(h_{FA}) \\ z(h_H) - z(h_{FA}) &= \hat{d}' \end{aligned} \tag{2.9}$$

Zur Vereinfachung der Notation sei an dieser Stelle vereinbart, dass $z(h_H) = z_H$ und $z(h_{FA}) = z_{FA}$ gelte und von nun an als Bezeichnung für diese Werte benutzt wird. Diese Art der Transformation einer Wahrscheinlichkeit p in ihren zugehörigen Wert z auf die Dimension, auf welcher die zugehörige Zufallsvariable verteilt ist, wird auch als Transformation in Gauss'sche Koordinaten bezeichnet. Es ergibt sich für die obigen

Gleichungen 2.8 und 2.9 die vereinfachte Schreibweise:

$$\hat{\lambda} = -z_{FA} \tag{2.10}$$

$$\hat{d'} = z_H - z_{FA} \tag{2.11}$$

Damit ist es möglich, sowohl die Lage der Signal-Verteilung d' auf der Dimension des subjektiven Eindrucks relativ zur *noise*-Verteilung und die Lage des Antwortkriteriums einer Versuchsperson λ für eine bestimmte experimentelle Situation zu bestimmen. In Abschnitt 2.2.1 wurde ein Beispielexperiment nach Wickens (2002) beschrieben, in welchem zwei unterschiedliche Instruktionen für die Versuchsperson zu verschiedenen *hit*- bzw. *false alarm*-Raten geführt haben. Dabei konnte davon ausgegangen werden, dass diese Veränderung nicht auf eine veränderte Sensitivität der Versuchsperson zurückzuführen ist, sondern sich deren Antwortkriterium verändert haben muss.

Führt man nun für die beiden beispielhaften Experimentalsitzungen die Parameterschätzung anhand obenstehenden Gleichungen durch, ergeben sich für beide Sitzungen tatsächlich fast gleiche Schätzungen für die Lage der Signal-Verteilung d', aber zwei recht unterschiedliche Kriterien λ_1 und λ_2 (s. Tabelle 2.3). Es stellt sich daher die Frage, wie diesem Umstand Rechnung getragen werden kann.

2.3.2 *Receiver Operating Characteristics*

Aus den obigen Ausführungen ist bereits klar, dass bei festem Verhältnis zwischen Signal- und *noise*-Verteilung, die *hit*- und die *false alarm*-Rate von der Lage des Antwortkriteriums λ abhängig sind und gemeinsam variieren. Um sich dies zu veranschaulichen, ist es hilfreich, die *hit*- und die *false alarm*-Rate der beiden Experimen-

Tabelle 2.3. Ergebnis der Parameterschätzung für das Beispielexperiment nach Wickens (2002).

Sitzung	h_H	h_{FA}	$\hat{d}'$	$\hat{\lambda}$
1	.82	.46	1.02	.10
2	.55	.19	1.00	.88

te gegeneinander in einem Diagramm abzutragen. Dabei ist es Konvention, die *false alarm*-Rate auf der Abszisse und die *hit*-Rate auf der Ordinate abzutragen. Dies ist in Abbildung 2.4 für das Beispielexperiment nach Wickens (2002) dargestellt.

Man kann sich nun vorstellen, dass die Versuchsperson nicht nur über zwei solcher Antwortkriterien verfügt, sondern über unendlich viele, welche sich über das gesamte Spektrum der Dimension des subjektiven Eindrucks hinweg verteilen. Um jedes einzelne dieser Antwortkriterien in einer festen experimentellen Situation zu aktivieren, ist es nur notwendig, die jeweils passenden Anreize für *hits*, *misses*, *false alarm* bzw. *correct rejections* zu finden, ähnlich, wie dies in Abschnitt 2.2.1 beschrieben ist. Man spricht in diesem Zusammenhang von sogenannten *pay-off*-Matrizen, welche jedem der vier Ereignisse einen nummerischen Wert zuweisen, um so eine Belohnung bzw. Bestrafung der Versuchsperson beim Auslösen des jeweiligen Ereignisses zu beschreiben. Die beiden *pay-off*-Matrizen für das Beispielexperiment aus Abschnitt 2.2.1 sind in Tabelle 2.4 dargestellt.

Führte man nun hypothetisch das gleiche Experiment mit der gleichen Versuchsperson für jede dieser unendlich vielen *pay-off*-Matrizen durch, so würde die Versuchsperson für jede Matrix ein anderes Antwortkriterium von sehr liberal, also weit links auf der Dimension des subjektiven Reizeindrucks, bis sehr konservativ, also weit rechts auf dieser Dimension, wählen. Dies führte jeweils zu einem bestimmten Paar aus *hit*- und *false alarm*-Rate für jede einzelne Matrix. Diese unendlich vielen Paare

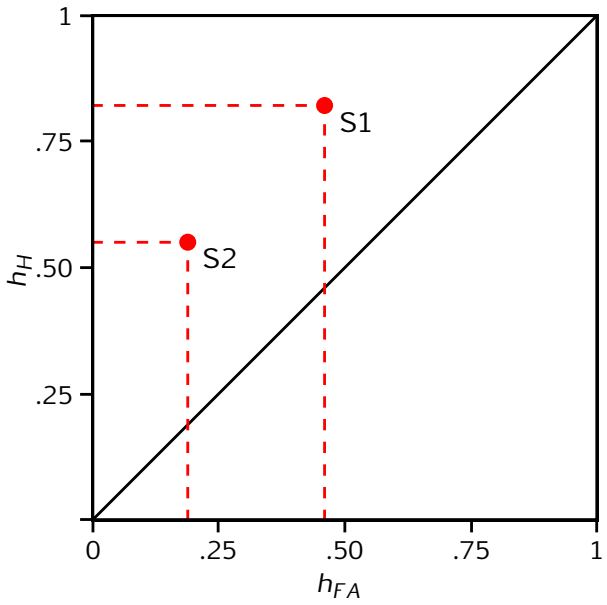

Abbildung 2.4. Plot der empirischen *hit*- und *false alarm*-Raten aus der ersten (S1) und zweiten (S2) Sitzung des Beispielexperiments nach Wickens (2002).

ließen sich alle gemeinsam in einer Abbildung gegeneinander abtragen und man erhielte als Ergebnis unendlich viele Punkte, welche sich alle auf einer gemeinsamen gebogenen Linie befänden. Diese gebogene Linie wird aufgrund des Ursprungs der Signalentdeckungstheorie im Radarwesen (vgl. Abschnitt 2.1.1) oft als *receiver operating characteristic*, bzw. kürzer *ROC* oder *ROC*-Kurve bezeichnet. Gebräuchlich ist auch der neutralere Begriff Iso-Sensitivitätskurve, welcher sich aus der Tatsache herleitet, dass sich aus dieser Kurve für eine gleichbleibende Sensitivität d' einer Versuchsperson für alle denkbaren Kriterien λ die zugehörige *hit*- und *false alarm*-Rate ablesen lässt. Für das Beispielexperiment nach Wickens (2002) ist die *ROC*-Kurve unter der Annahme von $d' = 1$, was der Schätzung der Sensitivität in der zweiten Sitzung entspricht, in Abbildung 2.5 dargestellt. An dieser Stelle ist es wichtig, zu bemerken, dass die *ROC*-Kurve ein theoretisches Konstrukt und im *equal-variance*-Modell symmetrisch ist.

Tabelle 2.4. Übersicht über die *pay-off*-Matrizen, die in den beiden Sitzungen des Beispielexperiments nach Wickens (2002) verwendet wurden.

	Sitzung 1		Sitzung 2	
	„*S*“	„*N*“	„*S*“	„*N*“
S	10	0	0	0
N	0	0	0	10

Da der Lageparameter d', wie zu Beginn dieses Abschnitts bereits beschrieben, ein direktes Maß für die Sensitivität der Versuchsperson darstellt, soll hier nicht unerwähnt bleiben, welche Rolle d' für die Form von *ROC*-Kurven spielt. Aus den bisherigen Betrachtungen ist klar, dass die Größe der *hit*-Rate bei einem festen Kriterium λ im *equal-variance*-Modell nur von d' abhängt (vgl. auch Gleichung 2.7). Verschiebt sich die Signal-Verteilung nun also nach rechts auf der Dimension des subjektiven Eindrucks, ist dies gleichbedeutend mit einer höheren Sensitivität der Versuchsperson und einem größeren Wert für d'. Diese Verschiebung von d' führt gleichzeitig zu einer größeren *hit*-Rate, während die *false alarm*-Rate konstant bleibt.

Für die Bestimmung einer *ROC*-Kurve bedeutet dies, dass nun für jedes einzelne Kriterium λ die *hit*-Rate über einer bestimmten, unveränderten *false alarm*-Rate einen größeren Wert annimmt. Somit liegt nun also jeder einzelne Punkt der *ROC*-Kurve „weiter oben“ in der Abbildung, was dazu führt, dass sich die gesamte *ROC*-Kurve weiter von der Winkelhalbierenden zwischen den Hauptachsen des Koordinatenssystems entfernt. In Abbildung 2.6 sind für verschiedene Werte von d' beispielhaft deren zugehörige *ROC*-Kurven dargestellt.

Die „Höhe“ einer *ROC*-Kurve, also ihr maximaler Abstand von der Winkelhalbierenden, hängt demnach einzig von der relativen Lage der Signal-Verteilung zur *noise*-Verteilung, also dem Parameter d', ab. Dieser lässt sich aus einem einzigen Paar aus

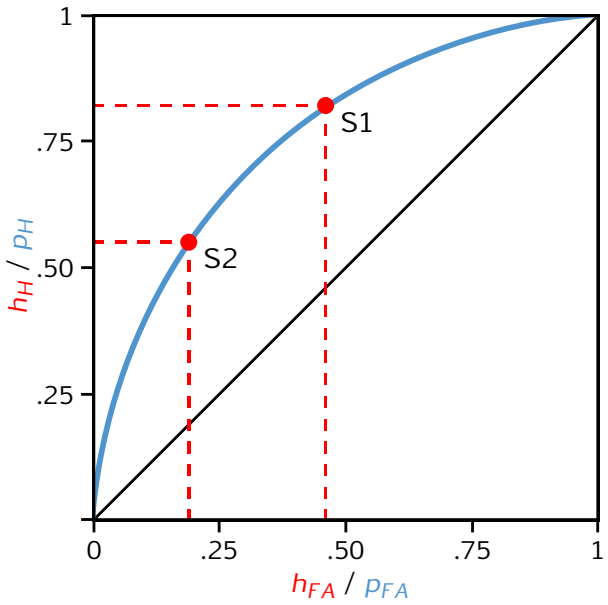

Abbildung 2.5. Plot der empirischen *hit*- und *false alarm*-Raten (rot) aus der ersten (S1) und zweiten (S2) Sitzung des Beispielexperiments nach Wickens (2002) auf der theoretischen *receiver operating characteristic* (*ROC*, blau) des *equal-variance*-Modells mit $d' = 1$.

hit- und *false alarm*-Rate bestimmen. Daher determiniert schon die Durchführung eines Experiments mit einer einzigen *pay-off*-Matrix die *ROC*-Kurve und es ist nicht nötig, das Experiment mit weiteren *pay-off*-Matrizen zu wiederholen. Gleichermaßen lassen sich nach der Bestimmung von d' in diesem einen Experiment für beliebige Kriterien jeweils die zu erwartenden *hit*- und *false alarm*-Raten mittels der Gleichungen 2.6 und 2.7 vorhersagen.

2.4 Das *unequal-variance*-Modell

Die Reduktion des Signalentdeckungsmodells um den Varianzparameter der Signal-Verteilung hat zwei Vorteile. So ist es einerseits möglich, aus nur zwei empirisch beobachteten Daten die restlichen freien Parameter zu schätzen und andererseits kann diese Parameterschätzung mittels zweier einfacher Gleichungen beschrieben werden. Allerdings kann im Allgemeinen nicht davon ausgegangen werden, dass die Infor-

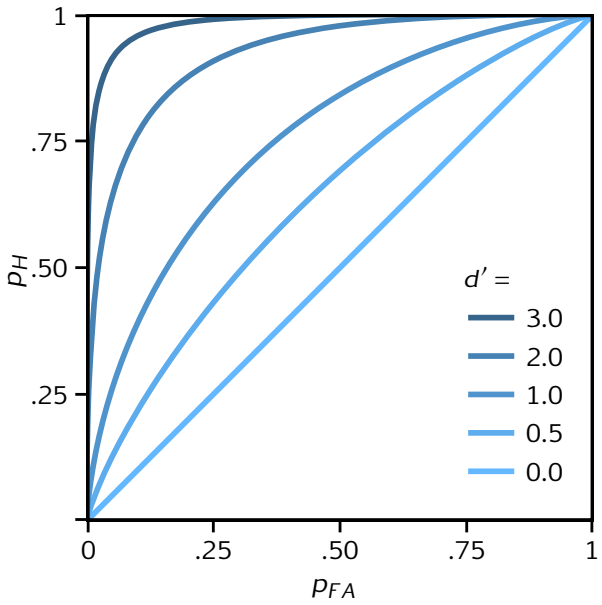

Abbildung 2.6. Plot mehrerer *ROC*-Kurven im *equal-variance*-Modell mit verschiedenen Werten für d'.

mation in Durchgängen mit Rauschen in der gleichen Weise variiert, wie in Durchgängen, in denen das Signal gemeinsam mit dem Rauschen präsentiert wird. Hier kommt zur Variabilität des Rauschens zusätzlich die Variabilität des Signals hinzu, so dass die Varianz in diesen Durchgängen o.B.d.A. größer ist. Es ist daher notwendig, auch die Varianz der Signal-Verteilung, und damit insgesamt die drei Parameter μ_s, σ_s^2 und λ, zu schätzen. Dies ist jedoch nicht mehr allein mittels eines einzigen Paares aus *hit*- und *false alarm*-Rate möglich.

2.4.1 Parameterschätzung

Angenommen, man habe nur ein einziges Signalentdeckungsexperiment durchgeführt. Dann ist es weiterhin, wie bereits in Abschnitt 2.3.1 dargestellt, möglich, die Schätzung für das Kriterium λ vorzunehmen, da hierfür bei Kenntnis der *noise*-Verteilung stets die empirisch bestimmte *false alarm*-Rate ausreichend ist. Da die *noise*-Verteilung auch im *unequal-variance*-Modell weiterhin standardnormalverteilt ist,

gilt somit entsprechend Gleichung 2.10 weiterhin, welche sich aus der Beziehung in Gleichung 2.3 herleitet, also:

$$\hat{\lambda} = -z_{FA} \tag{2.12}$$

Für die beiden anderen zu schätzenden Parameter μ_s und σ_s^2 lässt sich mit der Beziehung aus Gleichung 2.4 in ähnlicher Weise, wie dies in Gleichung 2.9 für das *equal-variance*-Modell geschehen ist, eine Gleichung für das *unequal-variance*-Modell bestimmen, die diese beiden Parameter enthält. Es gilt:

$$\begin{aligned} h_H &= 1 - \Phi\left(\frac{\hat{\lambda} - \hat{\mu}_s}{\hat{\sigma}_s}\right) \\ &= \Phi\left(\frac{-\hat{\lambda} + \hat{\mu}_s}{\hat{\sigma}_s}\right) = \Phi\left(\frac{\hat{\mu}_s - \hat{\lambda}}{\hat{\sigma}_s}\right) \\ z_H &= \frac{\hat{\mu}_s - \hat{\lambda}}{\hat{\sigma}_s} \\ z_H &= \frac{\hat{\mu}_s - (-z_{FA})}{\hat{\sigma}_s} = \frac{\hat{\mu}_s + z_{FA}}{\hat{\sigma}_s} \\ z_H &= \frac{1}{\hat{\sigma}_s} z_{FA} + \frac{\hat{\mu}_s}{\hat{\sigma}_s} \end{aligned} \tag{2.13}$$

Wie man leicht sieht, handelt es sich bei Gleichung 2.13 um eine lineare Funktion, welche jedoch zwei Unbekannte beinhaltet und daher nicht aufgelöst werden kann. Die Schätzung der beiden Parameter μ_s und σ_s^2 ist daher mit nur einer *hit*-Rate nicht möglich.

Rückblickend lässt sich auch Gleichung 2.11 zur Bestimmung von d' als Ergebnis der Auflösung einer solchen linearen Funktion nach d' auffassen. Diese linearen Funktionen lassen sich als Geraden in einem Koordinatensystem wie in Abbildung 2.7 darstellen. Dabei befindet sich z_{FA} wie schon dessen Pendant p_{FA} auf der Abszisse und z_H ebenfalls wie p_H auf der Ordinate. Man spricht hierbei von einer Darstellung der *hit*- und *false alarm*-Raten in Gauss'schen Koordinaten (Wickens, 2002).

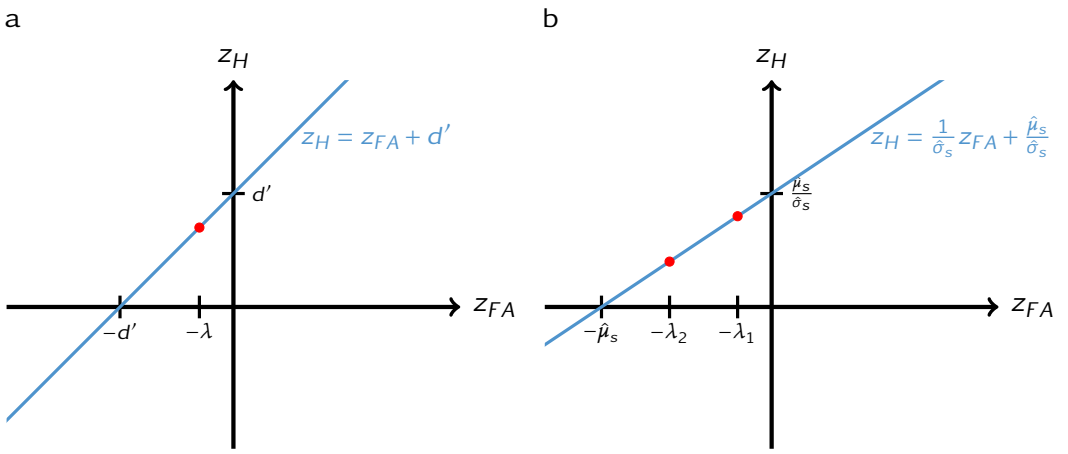

Abbildung 2.7. Plot der Iso-Sensitivitätskurven bzw. *receiver operating characteristics* (*ROC*-Kurven) im *equal-variance*-Modell (a) und *unequal-variance*-Modell (b) in Gauss'schen Koordinaten, nebst jeweiliger linearer Funktion mit zugehörigen Parametern.

Aus der Abbildung 2.7a erkennt man, dass sich der Parameter d' des *equal-variance*-Modells grafisch leicht bestimmen lässt, indem eine Gerade im Winkel von 45° zu den Hauptachsen des Koordinatensystems soweit verschoben wird, bis diese den Punkt schneidet, welcher durch ein einziges Paar aus *hit*- und *false alarm*-Rate vorgegeben ist. Dies folgt aus der linearen Funktion

$$z_H = z_{FA} + d' \tag{2.14}$$

welche sich aus Gleichung 2.11 zur Bestimmung von d' herleitet. Wie man leicht sieht, lassen sich die Nullstellen auf den jeweiligen Achsen x_0 und y_0 bestimmen, indem jeweils für z_H bzw. z_{FA} Null eingesetzt wird:

$$\begin{aligned} z_H &= z_{FA} + d' & z_H &= z_{FA} + d' \\ 0 &= x_0 + d' & y_0 &= 0 + d' \\ x_0 &= -d' & y_0 &= d' \end{aligned} \tag{2.15}$$

Auf ähnliche Weise lassen sich nun auch die Parameter der linearen Funktion des *unequal-variance*-Modells bestimmen (vgl. Abbildung 2.7b). Da dessen Gerade allerdings o.B.d.A. nicht im 45°-Winkel zu den Hauptachsen verläuft, sondern beliebig variieren kann, sind aus einfachen geometrischen Gründen zwei Punkte im Koordinatensystem, also letztlich zwei *hit-/false alarm*-Raten-Paare notwendig, um deren Lage zu determinieren. Diese beiden Paare erhält man, indem man ein Experiment mit der gleichen Versuchsperson aber zwei unterschiedlichen *pay-off*-Matrizen wiederholt und diese so dazu bewegt, ihr Kriterium λ in den beiden Sitzungen zu verändern (s. dazu auch die Abschnitte 2.2.1 und 2.3.2).

Für die Nullstellen der linearen Funktion des *unequal-variance*-Modells ergeben sich:

$$\begin{aligned} z_H &= \frac{1}{\hat{\sigma}_s} z_{FA} + \frac{\hat{\mu}_s}{\hat{\sigma}_s} \\ 0 &= \frac{1}{\hat{\sigma}_s} x_0 + \frac{\hat{\mu}_s}{\hat{\sigma}_s} \\ -\frac{\hat{\mu}_s}{\hat{\sigma}_s} &= \frac{1}{\hat{\sigma}_s} x_0 \\ x_0 &= -\hat{\mu}_s \end{aligned} \qquad \begin{aligned} z_H &= \frac{1}{\hat{\sigma}_s} z_{FA} + \frac{\hat{\mu}_s}{\hat{\sigma}_s} \\ y_0 &= \frac{1}{\hat{\sigma}_s} * 0 + \frac{\hat{\mu}_s}{\hat{\sigma}_s} \\ y_0 &= \frac{\hat{\mu}_s}{\hat{\sigma}_s} \end{aligned} \tag{2.16}$$

Damit ist klar, dass sich nach der Bestimmung der Lage der linearen Funktion im Koordinatensystem, die zu schätzenden Parameter für das *unequal-variance*-Modell leicht aus Nullstellen der Funktion, welche im Koordinatensystem abgelesen werden können, auf folgende Weise berechnen lassen:

$$\hat{\mu}_s = -x_0 \tag{2.17}$$

und

$$\begin{aligned} y_0 &= \frac{\hat{\mu}_s}{\hat{\sigma}_s} \\ y_0 &= \frac{-x_0}{\hat{\sigma}_s} \\ \hat{\sigma}_s &= \frac{-x_0}{y_0} \end{aligned} \tag{2.18}$$

Es ist jedoch wünschenswert, die Parameter der linearen Funktion nicht nur grafisch ermitteln zu können, da es ohne Weiteres möglich ist, mehr als zwei *hit*-/*false alarm*-Paare durch vielmaliges Durchführen desselben Experiments mit unterschiedlichen *pay-off*-Matrizen zu gewinnen und diese durch die zufälligen Schwankungen von Daten aus Experimenten vermutlich nicht alle auf ein und derselben Gerade lokalisiert sind[4]. In diesem Fall ist es nötig, eine Gerade zu finden, welche die Daten bestmöglich repräsentiert. Dieses Problem lässt sich auf grafischem Wege nur unzureichend lösen. Es ist daher eine algebraische Lösung notwendig.

Der Gedanke, eine lineare Regression der Datenpunkte durchzuführen und so die Parameter als jene der Regressionsgleichung zu bestimmen, ist naheliegend, stellt aber einen Trugschluss dar. Im Regressionsmodell wird die Prädiktorvariable x für die Kriteriumsvariable y als ein exakter Wert angenommen, welcher frei von Varianz ist (Hays, 1994; Maxwell & Delaney, 2004). Dies trifft jedoch im Signalentdeckungsmodell nicht zu, in welchem sowohl z_H als auch z_{FA} transformierte Zufallsvariablen sind und somit einer gewissen Varianz unterliegen (Wickens, 2002).

Es ist daher ein anderer Ansatz zur Bestimmung der Parameter notwendig. Da der wahrscheinlichkeitstheoretische Unterbau des hier verwendeten Signalentdeckungs-

[4] Darüber hinaus ist es aus experimentellen und statistischen Gründen absolut wünschenswert, nicht nur über zwei Datenpunkte zu verfügen, da sich für zwei beliebige Punkte immer eine perfekte Anpassung herstellen lässt, jedoch keine Freiheitsgrade für einen Anpassungstest verbleiben.

modells ohne Weiteres formal beschreibbar ist, steht hierfür das *maximum-likelihood*-Verfahren zur Verfügung. Dabei wird ausgehend von groben, möglicherweise nicht sehr guten Schätzungen der Parameter mittels eines iterativen Verfahrens versucht, diese Schätzungen Durchlauf für Durchlauf zu verändern, um die Anpassung der geschätzten Geraden an die Datenpunkte zu verbessern. Sobald die Verbesserung der Anpassung von einem Durchgang zum nächsten ein bestimmtes, beliebig klein gewähltes Kriterium unterschreitet, werden die in diesem Durchgang geschätzten Parameter akzeptiert.

Obwohl bzw. gerade weil es sich bei dem *maximum-likelihood*-Verfahren um ein Standardverfahren bei der Parameterschätzung handelt, wird die tatsächliche Durchführung der Parameterschätzung oft einem Computer überlassen, da dieser wesentlich schneller und fehlerfrei dazu in der Lage ist, viele Anpassungsdurchläufe in kurzer Zeit vorzunehmen. Für die Schätzungen von Signalentdeckungsparametern in der vorliegenden Arbeit wurde zu diesem Zweck die frei verfügbare Statistiksoftware R (R Core Team, 2015), speziell das Zusatzpaket *ordinal* (Christensen, 2015b), verwendet. Das Paket *ordinal* schätzt dabei die Parameter unter Verwendung eines *cumulative link model*-Ansatzes (Agresti, 2002; Christensen, 2015a). Ein Abdruck des Scripts zur Aufbereitung eines Datensatzes und der Parameterschätzung findet sich zur Information in Anhang A).

2.4.2 Verlagerung des Kriteriums mittels *Rating*-Verfahren

Während es in vielen experimentellen Kontexten nur aufgrund des Faktors Zeit praktische Grenzen bei der Wiederholung des Experiments mit immer neuen *pay-off*-Matrizen zur Verlagerung des Antwortkriteriums λ gibt, ist dies unter einem anwendungsorientierten Ansatz nicht ohne Weiteres möglich. Hier ist es wünschenswert, schon

nach der Durchführung einer einzigen Sitzung die Leistung der Versuchsperson bestimmen zu können. Möchte man aus theoretischen oder inhaltlichen Gründen nicht auf das *equal-variance*-Modell zurückgreifen, ist es zur Parameterschätzung jedoch zwingend erforderlich, über Daten mit mindestens zwei, besser drei[5], zugrundeliegenden Kriterien zu verfügen. Hier schafft das *Rating*-Verfahren Abhilfe.

Das bisher beschriebene theoretische Signalentdeckungsmodell der beiden normalverteilten Zufallsvariablen X_s bzw. X_n, welche jeweils bei der Präsentation von Signal bzw. *noise* realisiert werden, bleibt hierbei unberührt. Statt aber das Experiment nun immer wieder neu mit zwar unterschiedlichen *pay-off*-Matrizen aber den immer gleichen Antwortkategorien „Signal anwesend“ bzw. „Signal abwesend“ durchzuführen, wird das Experiment nur ein einziges Mal durchgeführt, die Versuchsperson erhält jedoch die Möglichkeit, ihre Antwort mittels eines Sicherheits*ratings* abzustufen.

Für dieses *Rating* müssen mindestens drei Stufen bzw. Antwortkategorien zur Verfügung stehen, so dass sich später die mindestens notwendigen zwei Datenpunkte zur Determinierung der Lage einer Geraden im Gauss'schen Koordinatensystem bestimmen lassen. Die Kategorien können dabei beliebig benannt werden, z.B. von „sicher *noise*“ über „eher *noise*“ und „eher Signal“ bis hin zu „sicher Signal“ (s. z.B. McNicol, 2005, S. 25 oder Egan et al., 1959). Eine beispielhafte Datentabelle für dieses vierstufige *Rating* ist im oberen Teil von Tabelle 2.5 dargestellt.

Die maximale Kategorienanzahl ist theoretisch unbegrenzt, in der Praxis beschränkt man sich jedoch üblicherweise auf vier bis zehn Kategorien, da Versuchspersonen diese konsistent benutzen können müssen (McNicol, 2005). Dabei ist es wichtig, festzuhalten, dass die üblicherweise durch den Versuchsleiter vorgenommene Benennung der Kategorien einerseits für jede Versuchsperson unterschiedlich auf der Dimension

[5] s. Fußnote 4 bezüglich der Anpassung einer Geraden an zwei Datenpunkte.

Tabelle 2.5. Beispielhafte Datentabelle für ein *Rating*-Verfahren im *unequal-variance*-Modell. Im oberen Teil finden sich bedingte absolute Häufigkeiten, im unteren Teil ausgehend von der größten Signalausprägung, in diesem Falle also von rechts, aufsummierte bedingte relative Häufigkeiten (ohne Korrektur für Extremwerte, s. Abschnitt 2.6).

	Rating				
	„1“: „sicher *noise*“	„2“: „eher *noise*“	„3“: „eher Signal“	„4“: „sicher Signal“	Summe
	absolute Häufigkeiten				
S	5	10	25	60	100
N	65	15	10	10	100
	aufsummierte bedingte relative Häufigkeiten				
S	1	.95	.85	.60	
N	1	.35	.20	.10	

des subjektiven Eindrucks lokalisiert ist und andererseits die Semantik der Benennung im Signalentdeckungsmodell keinerlei Repräsentation hat. Dort ist einzig die Reihenfolge der Kategorien entscheidend, wobei große Werte eher für das Vorliegen des Signals sprechen sollen.

Die Versuchsperson kann auf diese Weise bei jeder Reizpräsentation selbst entscheiden, wie sicher sie sich ist, ein Signal wahrgenommen oder nicht wahrgenommen zu haben und die entsprechende Kategorie auswählen. So kann sie für die verschiedenen Antwortkategorien jeweils unterschiedliche Kriterien auf ihrer Dimension des subjektiven Eindrucks von eher liberal bis eher konservativ festlegen. Dies ist beispielhaft für ein vierstufiges *Rating* in Abbildung 2.8 dargestellt, wobei hier die Kategorien neutral mit Zahlen benannt sind, diese jedoch ohne Weiteres die etwas weiter oben im Text genannten Namen tragen könnten.

Um nun die Datenpunkte in Gauss'schen Koordinaten für die Bestimmung der Geraden zu ermitteln, wird iterativ vorgegangen: Für jedes Kriterium, also jede Grenze

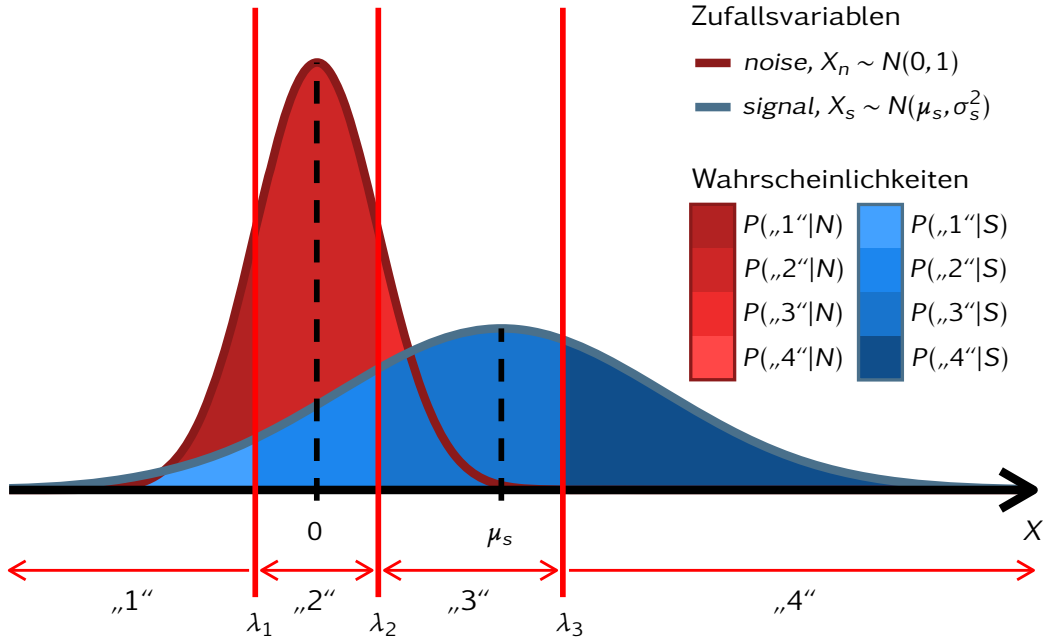

Abbildung 2.8. Dargestellt sind die *noise*- und Signal-Verteilung eines *unequal-variance*-Signalentdeckungsmodells beim *Rating*-Verfahren mit den vier Antwortkategorien „1“ bis „4“ und den sich als Flächen unter den jeweiligen Verteilungen ergebenden bedingten Wahrscheinlichkeiten.

zwischen zwei Kategorien, wird ausgehend von der größten Signalausprägung das Experiment als eine einfache *Yes*/*No*-Aufgabe behandelt und jeweils ein Paar aus *hit*- und *false alarm*-Rate bestimmt. Dazu werden die jeweiligen bedingten relativen Häufigkeiten rechts vom aktuell behandelten Kriterium aufsummiert. Es ergeben sich somit von rechts nach links ansteigende Paare vn *hit*- und *false alarm*-Raten, welche für das oben genannte Beispiel im unteren Teil von Tabelle 2.5 berechnet sind. Mittels der Umkehrfunktion der kumulativen Verteilungsfunktion $z(p)$ werden diese Paare jeweils in Gauss'sche Koordinaten umgerechnet und anschließend als Datenpunkte im Gauss'schen Koordinatensystem abgetragen. Auf Grundlage dieser Datenpunkte erfolgt nun die Anpassung der bestmöglichen Geraden mittels des *maximum-likelihood*-Verfahrens, aus deren Anstieg und Achsenabschnitt sich anschließend μ_s

und σ_s^2 auf die in Abschnitt 2.4.1 beschriebene Art und Weise berechnen lassen.

Vorteilhaft bei der Verwendung des *Rating*-Verfahrens ist, dass das Experiment nur ein einziges Mal durchgeführt werden muss, um dennoch ausreichend viele Datenpunkte zur Schätzung der Parameter für ein *unequal-variance*-Modell zu erhalten. Hier liegt jedoch gleichzeitig der größte Nachteil, da sich die gleiche Menge an Antworten einer Versuchsperson in den einzelnen Präsentationen von Signal- und *noise*-Durchgängen nun auf mehrere Kategorien aufteilt und so ggf. das Gesetz der großen Zahl (vgl. Abschnitt 2.3.1) unterlaufen wird und die Schätzungen der Auftretenswahrscheinlichkeiten der einzelnen Ereignisse nicht mehr die gleiche Stabilität aufweisen.

2.4.3 *Receiver Operating Characteristics*

Auch im *unequal-variance*-Modell lässt sich eine *ROC*-Kurve bestimmen, indem man bei bekannter Lage μ_s und Varianz σ_s^2 der Signal-Verteilung hypothetisch für jedes denkbare Kriterium auf dem gesamten Spektrum der Dimension des subjektiven Eindrucks jeweils die *hit*- bzw. *false alarm*-Rate bestimmt und diese gegeneinander auf gleiche Weise wie im *equal-variance*-Modell in einem Koordinatenssystem abträgt (vgl. Abschnitt 2.3.2). Für den allgemeinen Fall, dass die Varianzen von Signal- und *noise*-Verteilung verschieden sind, ist die *ROC*-Kurve jedoch nicht mehr symmetrisch. In Abbildung 2.9 sind mehrere beispielhafte *ROC*-Kurven mit unterschiedlichen Werten für μ_s und immer gleicher Varianz $\sigma_s^2 = 7$ dargestellt. Die *ROC*-Kurve mit dem Parameter $\mu_s = 3$ entspricht dabei der Lage und Varianz der Signal-Verteilung in den Abbildungen 2.2 und 2.8.

Einen einfacheren Weg, *ROC*-Kurven zu bestimmen, bietet sich durch die Kenntnis der Iso-Sensitivitätskurve in ihrer Darstellung in Gauss'schen Koordinaten. Aufgrund deren Linearität ist es dort sehr einfach, aus einer Reihe vorgegebener *false alarm*-

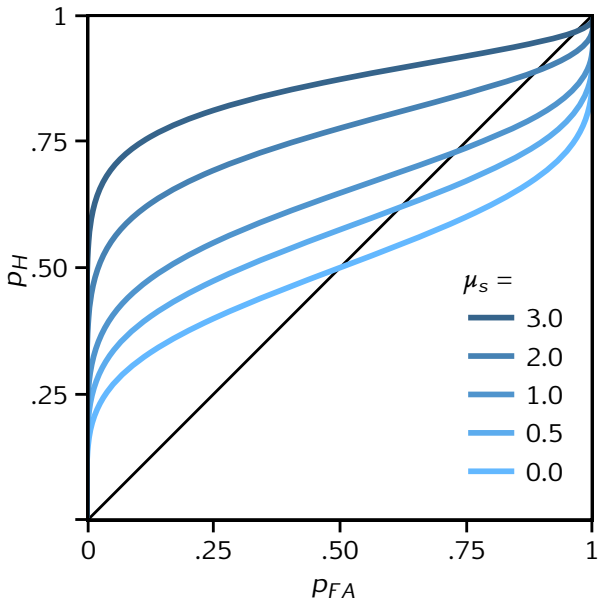

Abbildung 2.9. Plot mehrerer *ROC*-Kurven im *unequal-variance*-Modell mit verschiedenen Werten für μ_s und $\sigma_s^2 = 7$. Die *ROC* mit dem Parameter $\mu_s = 3$ entstammt der Lage und Überlappung der Verteilungen, wie sie in Abbildung 2.2 dargestellt ist.

Raten die zugehörigen *hit*-Raten schnell zu berechnen. Diese lassen sich anschließend leicht in die typischere Darstellung in Wahrscheinlichkeitskoordinaten zurücktransformieren und als Punkte in einem Koordinatenssystem mit Wahrscheinlichkeitskoordinaten eintragen.

Auf ebendiese Weise wurden alle in dieser Arbeit abgebildeten *ROC*-Kurven, ebenfalls unter Verwendung von R (R Core Team, 2015), speziell mithilfe des Zusatzpakets *sensR* (Christensen & Brockhoff, 2015), erstellt. *SensR* erlaubt es, auf einfache Weise eine große Menge von *hit*-/*false alarm*-Raten-Paaren mittels der im vorherigen Absatz beschriebenen Methode für gegebene Parameter μ_s bzw. σ_s zu berechnen. Wird eine ausreichend große Menge, in dieser Arbeit meist $n = 1000$, dieser Paare als Punkte in ein Koordinatensystem eingetragen und mittels Linien verbunden, entsteht der Eindruck einer durchgehenden, geschwungenen Kurve, welche die tatsächliche *ROC*-Kurve in sehr eindrücklicher Weise approximiert und für dieser Art Abbildungen völlig

ausreichend ist. Dies gilt gleichermaßen auch für das *equal-variance*-Modell.

2.5 Die Fläche unter der *ROC*-Kurve

Bei der Betrachtung der *ROC*-Kurve in Abbildung 2.9 wird klar, dass der maximale Abstand der Kurve von der Winkelhalbierenden der Hauptachsen des Koordinatensystems im *unequal-variance*-Modell kein vernünftig interpretierbares Maß für die Sensitivität einer Versuchsperson ist, wie dies noch im *equal-variance*-Modell der Fall war. In beiden Modellen zeigt sich jedoch, dass die *ROC*-Kurve umso weiter in Richtung der oberen Begrenzungslinie liegt, desto größer die Sensitivität der Versuchsperson ist. Hieraus ergibt sich die Überlegung, die sich unter einer *ROC*-Kurve befindliche Fläche als ein neues Maß für die Sensitivität einer Versuchsperson heranzuziehen. Diese Fläche unter der Kurve wird als *AUC* von engl. *area under curve* bezeichnet und besitzt interessante Eigenschaften: Im *equal-variance*-Modell nimmt die *AUC* genau dann den Wert ½ an, wenn $d' = 0$ ist, die Leistung der Versuchsperson also gerade auf Rateniveau liegt. Umso größer d' wird, also, umso weiter die Signalkurve nach rechts verschoben wird, desto größer wird auch die *AUC*, bis diese sich schließlich ihrem Grenzwert 1 annähert, wenn d' Richtung $+\infty$ strebt. Im *unequal-variance*-Modell nimmt die *AUC* gleichermaßen mit größeren Werten für μ_s bzw. kleineren Werten für σ_s^2 zu.

Die *AUC* lässt sich, wie die Flächen für die vier Ereignisse *hits*, *misses*, *false alarms* und *correct rejections* unter den Dichtefunktionen der beiden Normalverteilungen, als das Integral unter der *ROC*-Kurve bestimmen. Hierfür ist lediglich die Kenntnis der Funktion vonnöten, welche die *ROC* beschreibt. Angenommen diese Funktion sei $R(p)$, dann beschreibt diese eine beliebige *ROC* als $p_H = R(p_{FA})$ und es gilt für die

zugehörige *AUC*:

$$AUC = \int_0^1 R(p)dp \tag{2.19}$$

Die Funktion $R(p)$ lässt sich nun für das Signalentdeckungsmodell unter der Annahme der Normalverteiltheit der beiden Zufallsvariablen X_n und X_s bestimmen. Dies führt im Rahmen dieser Arbeit jedoch zu weit und daher sei an dieser Stelle auf die einschlägige Literatur verwiesen (z.B. Wickens, 2002). Dort ist nachzulesen, dass sich die *AUC* ergibt als:

$$AUC = \Phi\left(\frac{\mu_s}{\sqrt{1+\sigma_s^2}}\right) \tag{2.20}$$

2.6 Umgang mit „extremen“ Kategorien

In bestimmten Fällen, gerade wenn es sich um Datensätze aus Experimenten mit wenigen Durchgängen oder mit vielen Antwortkategorien handelt, kann es vorkommen, dass die Versuchsperson bei ihren Antworten eine oder mehrere dieser Kategorien ausschließlich bzw. überhaupt nicht benutzt. Diese „extremen“ Kategorien führen zu Problemen bei der Parameterschätzung.

Dies zeigt sich offensichtlich im *equal-variance*-Modell: Abhängig von der Signalstärke, der tatsächlichen Sensitivität der Versuchsperson oder bestimmten, extremen *pay-off*-Matrizen kann es vorkommen, dass die Versuchsperson in allen Signal-Durchgängen mit „Signal anwesend“ und in allen *noise*-Durchgängen mit „Signal abwesend“ antwortet. Somit ergeben sich für die *hit*- bzw. *false alarm*-Rate die Werte $h_H = 1$ und $h_{FA} = 0$. Für die Werte 0 und 1 betragen die Funktionswerte der Umkehrfunktion der kumulativen Verteilungsfunktion der Normalverteilung $z(p)$ jedoch $-\infty$

bzw. $+\infty$, so dass z_H und z_{FA} zwar bestimmt werden können, die auf diese Weise anhand der beiden Gleichungen 2.10 und 2.11 geschätzten Parameter $\hat{\lambda}$ und $\hat{d}'$ jedoch schwerlich interpretierbar sind.

Zur Lösung dieses Problems wurden verschiedene Korrekturverfahren vorgeschlagen (Hautus, 1995; Hautus & Lee, 1998; Stanislaw & Todorov, 1999; Wickens, 2002), welche letztlich alle darauf abzielen, den absoluten Häufigkeitswert einer oder mehrerer extremer Antwortkategorien mittels Hinzuaddierens oder Abziehens eines arbiträren Wertes zu korrigieren. All diesen Möglichkeiten ist gemein, dass hierdurch zwar einerseits eine eindeutige nummerische Lösung für die zu schätzenden Parameter gewonnen wird, diese jedoch stark von der tatsächlichen Korrekturmethode und der zugrundeliegenden Fallzahl abhängig ist. Hautus und Lee (1998) konnten zeigen, dass dabei die sogenannte log-lineare Korrektur die geeignetste Methode darstellt, da sie sich am günstigsten verhält.

Bei dieser Korrekturmethode wird zu allen bedingten absoluten Häufigkeitswerten der vier möglichen Ereignisse jeweils der Wert ½ addiert und anschließend die bedingten Randhäufigkeiten neu bestimmt. Demnach ergeben sich nun für die beiden zur Parameterschätzung herangezogenen bedingten relativen Häufigkeiten:

$$\begin{aligned} h'_H &= \frac{H_H + 1/2}{H_S + 1} \\ h'_{FA} &= \frac{H_{FA} + 1/2}{H_N + 1} \end{aligned} \tag{2.21}$$

Im Allgemeinen wird diese Korrektur der bedingten absoluten Häufigkeitswerte für alle betrachteten Daten, unabhängig vom Vorliegen extremer Kategorien, durchgeführt, um die Vergleichbarkeit von Schätzungen unterschiedlichen Ursprungs zu gewährleisten. Daher wurden alle Daten, die den in dieser Arbeit berichteten Parameterschätzungen zugrunde liegen, mit der log-linearen Korrekturmethode korrigiert.

2.7 Anwendung der Signalentdeckungstheorie in anderen Bereichen

Bei der bisherigen Betrachtung der Signalentdeckungstheorie wurde stets von „Signal“ und „Rauschen“ in einem wahrnehmungspsychologischen Kontext gesprochen. Als Beispiele für das Signal wurden schwache, kaum wahrnehmbare auditive oder visuelle Reize angeführt, die vor einen verrauschten Hintergrund präsentiert werden. Zwar geht die Signalentdeckungstheorie tatsächlich auf ein solches Szenario zurück, ist darauf jedoch keineswegs beschränkt, sondern lässt sich im Gegenteil in fast allen Bereichen anwenden, in denen Personen Entscheidungen unter Unsicherheit treffen müssen. An dieser Stelle sollen beispielhaft zwei Einsatzbereiche dargestellt werden, einen Überblick über die zahlreichen anderen möglichen Anwendungsfelder gibt u.a. Hutchinson (1981).

2.7.1 Gedächtnispsychologie

So fand die Signalentdeckungstheorie zunächst Anwendung in der Gedächtnispsychologie, insbesondere bei der Auswertung von *recognition*-Experimenten (Banks, 1970; McNicol, 2005; Parks, 1966; Wickens, 2002). In diesen Experimenten ist es Aufgabe der Versuchsperson, aus einer Menge von einzeln nacheinander präsentierten Reizen, z.B. Bildern oder Wörtern, diejenigen Reize wiederzuerkennen, welche ihr bereits in einer vorangegangenen Phase des Experiments präsentiert worden sind. In solcherart Experimenten wird jedoch im Allgemeinen eine zu große Menge an Reizen präsentiert, als dass sich die Versuchsperson an alle perfekt erinnern könnte. Man geht daher davon aus, dass zwar jeder Reiz einen bestimmten subjektiven Eindruck der Vertrautheit erzeugt, dieser jedoch mehr oder weniger groß ausfallen kann, so

dass dieser subjektive Eindruck wiederum als die Zufallsvariable X aufgefasst werden kann.

Es lässt sich nun leicht argumentieren, dass es sich bei den bereits zuvor präsentierten, „alten“ Reizen bzw. bei den von ihnen erzeugten Gedächtnisspuren, welche ein bestimmtes Ausmaß an Vertrautheit bei der Versuchsperson hervorrufen (Abdi, 2007), um die Signale handelt. Gleichzeitig ist dann klar, dass die noch nicht zuvor präsentierten, „neuen“ Reize als das Rauschen aufgepasst werden können, da auch diese einen subjektiven Eindruck von Vertrautheit erzeugen können. Dieser sollte zwar in der Regel geringer sein als bei alten Reizen, ist jedoch, genau wie der subjektive Eindruck bei alten Reizen, zufälligen Schwankungen unterlegen, die zum Beispiel ihre Quelle in der allgemeinen neuronalen Aktivität des Gehirns oder anderen, alten Gedächtnisinhalten, welche nicht aus dem Experiment stammen, haben können (Abdi, 2007).

2.7.2 Medizin

In ähnlicher Weise wurde die Signalentdeckungstheorie auf andere Fachgebiete auch außerhalb der Psychologie adaptiert. So finden sich in der Medizin Beispiele, in denen die Güte von Diagnoseverfahren oder die Diagnoseentscheidungen von Ärzten mittels Signalentdeckungstheorie untersucht wurden (Altman, 1991; Beck, 1991; Bland, 2000; Lalkhen & McCluskey, 2008; Lusted, 1971a, 1971b; Parikh, Mathai, Parikh, Chandra Sekhar & Thomas, 2008; Sacks, Chalmers & Smith, 1983).

Hierbei wird das Diagnoseverfahren, z.B. ein Bluttest auf das Vorliegen einer bestimmten Erkrankung, bei einer Menge von Menschen angewendet und es stellt sich die Frage, wie oft dieses Verfahren die tatsächlich Kranken bzw. tatsächlich Gesunden als eben diese identifiziert, also keine Fehldiagnosen produziert werden. Gleich-

falls wurde oft untersucht, inwieweit Radiologen dazu in der Lage sind, aus einem Röntgenbild ablesen zu können, ob ein Patient tatsächlich an Krebs erkrankt ist oder nicht.

In diesen Beispielen entspricht jeweils das Vorliegen einer Erkrankung dem Signal, während die Blutwerte oder Röntgenbilder gesunder Menschen dem Rauschen entsprechen. Das Ziel, Fehldiagnosen zu vermeiden, entspricht dem Ziel, die Auftretenshäufigkeiten der aus der Signalentdeckungstheorie bekannten Ereignisse *misses* und *false alarms* möglichst gering, bzw. die Auftretenshäufigkeiten von *hits* und *correct rejections* möglichst hoch zu halten. Für die beiden letztgenannten haben sich in der Medizin die Begriffe Sensitivität und Spezifizität eingebürgert (Wickens, 2002).

2.8 Übertragung des Signalentdeckungsmodells auf Prüfungen

In ähnlicher Weise, wie im vorherigen Abschnitt beschrieben, lässt sich das Signalentdeckungsmodell nun auf den für die vorliegende Arbeit wichtigen Prüfungskontext mit *MR*-Aufgaben übertragen: Prüflinge sollen in einer Prüfung die richtigen von den falschen Alternativen unterscheiden und entsprechend den Anforderungen des Prüfungsformats kennzeichnen, in einem *MC*-Format also, indem sie richtige Alternativen markieren und falsche Alternativen freilassen. Jede einzelne Alternative stellt dabei die Präsentation eines Reizes dar. Es können nun im Sinne der Signalentdeckungstheorie, richtige Alternativen als Signale und falsche Alternativen als Rauschen aufgefasst werden. Die Prüflinge nehmen die Rolle der Versuchspersonen ein, deren Sensitivität bestimmt werden soll.

Liest ein Prüfling den Text einer Alternative und verarbeitet diesen im Kontext mit

der Aufgabenstellung, so löst jede dieser Alternativen einen mehr oder weniger starken subjektiven Eindruck der Bekanntheit bzw. Richtigkeit aus, ähnlich wie in einem *recognition*-Experiment. Dieser subjektive Eindruck stellt die Grundlage der Entscheidung „Kreuz“ bzw. „kein Kreuz“ bei dieser Alternative dar und kann als die Zufallsvariable X aufgefasst werden. Für diese Zufallsvariable X wird nun, aufgrund der Voraussetzungen des Signalentdeckungsmodells, wiederum angenommen, sie sei unter den Bedingungen Signal, also richtige Alternative, bzw. Rauschen, also falsche Alternative, mit o.B.d.A. unterschiedlichen Parametern normalverteilt. Dabei soll gelten, dass größere Werte für X, also ein größerer subjektiver Eindruck der Richtigkeit, eher für das Vorliegen des Signals, also einer tatsächlich richtigen Alternative, sprechen und demnach $\mu_s > \mu_n$ gelten soll.

Es lassen sich nun auf bekannte Weise nach der Prüfung für alle Prüflinge die *hit*- und *false alarm*-Raten bestimmen, wobei man in diesem Kontext neutraler von der *true positives*-Rate (*TPR*) und *false positives*-Rate (*FPR*) sprechen würde (Wickens, 2002). Aus diesen kann ein geeignetes Maß für die Sensitivität des Prüflings, z.B. die *AUC*, berechnet werden. Weiterhin ist es möglich, Prüfungen mit einem *Rating*-Verfahren durchzuführen, so dass Prüflinge für jede Alternative ihren subjektiven Eindruck der Richtigkeit, abgestuft nach ihren eigenen Kriterien der Sicherheit, bewerten können (s. Kapitel 1 und Abbildung 1.1). Eine Bestimmung der *AUC* ist mittels der Ausführungen in den Abschnitten 2.4.2 und 2.5 möglich.

3

Empirische Untersuchungen

Prüfungen in Pharmakologie in den Jahren 2012 bis 2014

3.1 Allgemeine Bemerkungen zur Methode

Als empirische Grundlage für diese Arbeit dienen drei Prüfungen, welche jeweils in den Jahren 2012, 2013 und 2014 stattfanden und in den nachfolgenden Abschnitten 3.2 bis 3.4 dargestellt sind. Da diese Prüfungen im Großen und Ganzen einem ähnlichen Muster folgen, soll dieses hier zunächst grob skizziert werden, um Wiederholungen zu vermeiden. An späterer Stelle wird nur auf die davon abweichenden Einzelheiten der jeweiligen Prüfung eingegangen.

Als Datenquelle dienten Prüfungen, welche Studierende im Rahmen des Moduls Pharmakologie und Toxikologie des Medizin-Studiengangs der Martin-Luther-Universität Halle-Wittenberg (MLU) unter der Leitung von Herrn Prof. Dr. Joachim Neumann

ablegten (s. Studienordnung des Medizinstudiengangs und deren Änderungsordnung, Martin-Luther-Universität Halle-Wittenberg, 2009, 2012). Das Modul besteht aus Vorlesungen und Seminaren und wird mittels einer obligatorischen und bestehensrelevanten Klausur jeweils zur Mitte (Zwischenklausur, ZK) und am Ende des Moduls (Abschlussklausur, AK) bewertet. Die Abschlussklausur findet in der letzten Semesterwoche statt und bildet den regulären Abschluss des Moduls. Im Rahmen dieser Abschlussklausur wurden den Prüflingen direkt im Anschluss an die regulären Aufgaben Zusatzaufgaben im *MR*-Format aus dem gerade geprüften Stoffgebiet gestellt. Diese Methode wurde gewählt, um Daten aus einem möglichst realitätsnahen Kontext zu erhalten, welche Rückschlüsse auf das tatsächliche Leistungsbild der Prüfling zulassen. Während die regulären Klausuren in allen drei Jahren in gleicher Weise im Einklang mit der allgemeingültigen Konvention in der Medizin als *SC*-Prüfung gestaltet und bewertet wurden, wurde das genaue Format der *MR*-Aufgaben in den verschiedenen Jahrgängen variiert und an die Erkenntnisse aus den vorangegangenen Jahren angepasst. Zur Auswertung für die Zusatzaufgaben wurden verschiedene *Scoring*-Methoden verwendet und darüber hinaus die Parameter eines Signalentdeckungsmodells für jeden Prüfling geschätzt. Als Maß für die Übereinstimmung der Leistungen in den regulären Klausuren und den Zusatzaufgaben im *MR*-Format wurden Korrelationen zwischen den jeweiligen Leistungsmaßen bestimmt.

3.1.1 Klausursituation

Die Zwischenklausur und die Abschlussklausur fanden jeweils in zwei etwa gleich großen, nebeneinander liegenden Hörsälen des Universitätsklinikums der MLU statt. Den Prüflingen wurde anhand von alphabetischen Teilnehmerlisten einer der beiden Hörsäle zugewiesen, so dass sich jeweils eine Hälfte der Prüflinge in jedem Hörsaal be-

fand. Nach dem Eintreffen der Prüflinge und der Überprüfung der Teilnahmeberechtigung an der Klausur wurden diese einzeln platziert.

Die Klausurbögen für die regulären Prüfungen lagen in vier Versionen mit jeweils zufälliger Reihenfolge der gleichen Aufgaben vor. Die Klausurbögen und in der Abschlussklausur zusätzlich die Bögen mit den Zusatzaufgaben wurden vor dem Eintreffen der Prüflinge im Hörsaal zufällig verteilt. Jeder Prüfling erhielt immer nur eine Variante der Klausurbögen und der Zusatzaufgaben. Alle Klausurbögen und Bögen mit Zusatzaufgaben wurden im Voraus mittels EvaSys bzw. EvaExam (Electric Paper Evaluationssysteme GmbH, 2015a, 2015b) in der zum Prüfungszeitpunkt aktuellen Version erstellt und die Antworten der Prüflinge im Anschluss an die Klausur mit Hilfe dieser Systeme automatisch elektronisch erfasst.

Während der gesamten Klausur waren in beiden Hörsälen jeweils mindestens zwei Lehrkräfte anwesend. Die Prüfung startete nach einer Belehrung zu Betrugsversuchen und dem Hinweis auf die zur Verfügung stehende Zeit und dauerte jeweils 75 Minuten. Davon waren 45 Minuten zur Beantwortung der regulären Klausuraufgaben vorgesehen und nach Ablauf dieser Zeit wurden die Klausurbögen eingesammelt. Den Prüflingen war es gestattet, den Raum auch vor Ablauf der Zeit nach der Abgabe aller Prüfungsbögen zu verlassen.

3.1.2 Klausuraufgaben

Die regulären Prüfungen zur Mitte und am Ende des Moduls bestanden jeweils aus 30 Aufgaben, welche für die jeweilige Prüfung anhand der Richtlinien von Case und Swanson, 2002 zur Erstellung von Typ-A-Aufgaben neu entworfen wurden. Es handelte sich demnach um Aufgaben im *SR*-Format mit fünf Antwortalternativen, wobei genau eine Alternative die richtige Lösung darstellte. Die Aufgaben für die Zwischen-

prüfung prüften Lehrstoff, welcher bis zu diesem Zeitpunkt in den Vorlesungen und Seminaren behandelt wurde, während die Aufgaben in der Abschlussprüfung Lehrstoff aus dem gesamten Modul, also auch bereits in der Zwischenklausur vorgekommenen Lehrstoff, prüften. Für die Beantwortung einer einzelnen Frage wurden nach allgemeinem Konsens in der Medizin 90 Sekunden veranschlagt, so dass sich eine Gesamtbearbeitungszeit von 45 Minuten für den regulären Prüfungsteil ergab.

Zur Benotung der Modulleistung wurden nur die Ergebnisse in den beiden regulären Prüfungen herangezogen. Jede einzelne Aufgabe wurde mit einem Punkt bewertet, sofern die vorgesehene richtige Alternative ausgewählt wurde. In jedem anderen Fall, wie kein Kreuz oder mehr als ein Kreuz in einer Aufgabe, wurden null Punkte für diese Aufgabe vergeben. Da jede Prüfung aus 30 gleichwertigen Aufgaben bestand, wurde zur Bewertung ein einfacher klassischer Summenwert der Punkte der einzelnen Aufgaben gebildet, so dass in jeder Prüfung maximal 30 Punkte erreicht werden konnten. Um das Modul zu bestehen, mussten mindestens 60% der Gesamtpunktzahl (GP), welche aus der Summe der Punkte aus der Zwischen- und der Abschlussklausur gebildet wurde, erreicht werden. Es waren also mindestens 36 korrekt beantwortete Aufgaben aus den 30 Zwischen- und den 30 Abschlussklausuraufgaben zum Bestehen notwendig, so dass es unumgänglich für die Prüflinge war, an beiden Prüfungen teilzunehmen.

3.1.3 Zusatzaufgaben

Es wurden in jedem Jahr jeweils zehn Zusatzaufgaben gestellt. Diese wurden aus Aufgaben abgeleitet, welche in früheren Jahren bereits als Typ-A-Aufgaben nach den Richtlinien von Case und Swanson, 2002 in Klausuren Verwendung fanden. Die Aufgaben wurden dahingehend ausgewählt, dass sie in früheren Prüfungen geeignete

SR	Welches der folgenden Antimykotika sollte am ehesten zur Therapie systemischer Pilzinfektionen verwendet?	
	×	Amphotericin B
		Flucytosin
		Ketoconazol
		Fluconazol
		Griseofulvin

MR	Welche der folgenden Antimykotika können zur Therapie systemischer Pilzinfektionen verwendet werden?	
	×	Amphotericin B
	×	Nystatin
	×	Tolnaftat
	×	Amorolfin
		Griseofulvin

Abbildung 3.1. Beispiel für die Anpassung einer *SR*-Aufgabe (oben) an das *MR*-Format mit vier richtigen Alternativen (unten). Alternativen, die mittels „×“ gekennzeichnet sind, stellen richtige Alternativen dar.

Werte für Schwierigkeit (zwischen .34 und .96) und Trennschärfe (zwischen .21 und .58) gezeigt hatten. In Abbildung 3.1 ist beispielhaft eine Aufgabe aus dem Jahr 2012 in ihrer ursprünglichen *SC*-Variante und in der angepassten *MR*-Variante dargestellt. In Anhang B findet sich eine Auflistung aller als Zusatzaufgaben verwendeten *MR*-Aufgaben, nach Jahren geordnet.

Um diese *SC*-Aufgaben in ein *MR*-Format zu überführen, wurden die Aufgabenstämme so wenig wie nötig angepasst, so dass aus grammatikalischen und/oder inhaltlichen Gründen nicht mehr nur eine, sondern auch mehrere richtige Alternativen möglich wären. Die Prüflinge wurden in einem Begleittext zu den Zusatzaufgaben darauf hingewiesen, dass entgegen ihrer bisherigen Prüfungserfahrung bei den Zusatzaufgaben immer die Möglichkeit besteht, dass keine, eine, zwei, drei, vier oder alle fünf

Alternativen richtig sein können. Es wurde betont, dass dennoch alle Aufgabenstämme aus Gründen der Übersichtlichkeit und Lesbarkeit im Plural formuliert sind, auch wenn aus inhaltlichen Gründen und entgegen der Grammatik keine oder eine Alternative richtig ist. Die Begleittexte der einzelnen Jahre sind in Anhang C abgedruckt.

Sofern es sich um eine Aufgabe mit genau einer richtigen Alternative handelte, wurden alle Alternativen aus dem *SR*-Format beibehalten. Für Aufgaben mit mehr als einer richtigen Alternative wurden entsprechend viele, nach dem Zufallsprinzip ausgewählte falsche Alternativen durch richtige Alternativen ersetzt. Aufgaben mit keiner richtigen Alternative wurden nicht verwendet, da es Prüflingen möglicherweise schwer fallen könnte, bei einem *MC*-Format bei einer ganzen Aufgabe kein einziges Kreuz zu setzen.

Die Beantwortung der Zusatzaufgaben war immer freiwillig und hatte keinen Einfluss auf die Bewertung der Prüfungsergebnisse in den regulären Prüfungen. Die genaue Ausgestaltung der Antwortschlüssel wurde in den verschiedenen Jahren verändert und an die Erkenntnisse aus dem Vorjahr angepasst. Für diesbezügliche Details sei an dieser Stelle daher auf die Abschnitte 3.2.2, 3.3.2 und 3.4.2 auf den nachfolgenden Seiten verwiesen.

3.1.4 Bestimmung der Bestehensgrenzen abhängig von der Ratewahrscheinlichkeit

In Abschnitt 1.4 wurden bereits einige Untersuchungen genannt, die das Ziel verfolgt haben, *MCQ*s zu verbessern, indem u.a. untersucht wurde, wie viele Alternativen vorgegeben werden sollten (Rodriguez, 2005). Grund dieser Untersuchungen ist die für Prüflinge stets vorhandene Möglichkeit, die richtige Antwort in einer Aufgabe zu erraten. Daher besteht für Prüflinge auch bei Nicht-Wissen weiterhin die Möglichkeit,

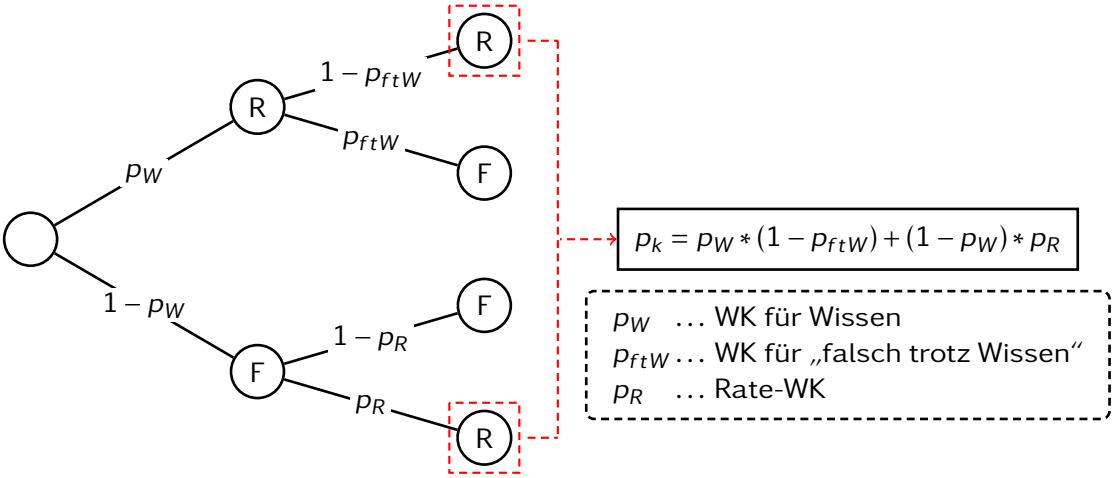

Abbildung 3.2. Wahrscheinlichkeitsbaum und Modellgleichung für das Modell nach Lukas (2015a, 2015b) für die Wahrscheinlichkeit (WK) einer korrekten Antwort p_k in einer Aufgabe. Eine richtige Antwort (R) ergibt sich aus dem obersten Pfad „Wissen, danach kein Schusselfehler“ und dem untersten Pfad „Nicht-Wissen, danach richtig geraten“.
Die beiden mittleren Pfade „Wissen, danach Schusselfehler“ und „Nicht-Wissen, danach falsch geraten“ führen zu falschen Antworten (F).

diese Aufgabe korrekt zu beantworten, wobei die Chance auf eine richtige Antwort dabei abhängig von der Größe der Ratewahrscheinlichkeit ist.

Da sich jedoch die Ratewahrscheinlichkeit für unterschiedliche Aufgabenformate ggf. stark unterscheiden kann und die erratenen Lösungen bei der Bewertung einer Prüfung nicht mit einbezogen werden sollen, ist eine entsprechende Korrektur vonnöten (Espinosa & Gardeazabal, 2010; Lord, 1975; Miles, 1973; Zimmerman & Williams, 1965, 2003). Zu diesem Zweck schlägt Lukas (2015a, 2015b) ein einfaches Wahrscheinlichkeitsmodell vor: Berechnet wird die Lösungswahrscheinlichkeit einer Aufgabe p_k in Abhängigkeit vom Wissen eines Prüflings p_W, unter Einbeziehung der konstanten Ratewahrscheinlichkeit des Aufgabenformats p_R und der konstanten Wahrscheinlichkeit p_{ftW} für sogenannte *careless errors* bzw. Flüchtigkeitsfehler, die zu falschen Antworten trotz Wissens führen. Das Modell lässt sich als Wahrscheinlichkeitsbaum, wie in Abbildung 3.2 dargestellt, beschreiben. Für die Wahrscheinlichkeit einer richtigen Antwort p_k ergibt sich damit eine lineare Funktion, die einzig vom Wissen

des Prüflings p_W abhängt, da die anderen beiden Parameter p_R und p_{ftW} für alle Aufgaben als konstant angenommen werden:

$$p_k = (1 - p_{ftW} - p_R) * p_W + p_R \tag{3.1}$$

Mithilfe dieser Modellgleichung ist es nun möglich, die Bestehensgrenze einer Prüfung zu berechnen. Üblicherweise wird hierfür das 50%-Kriterium ($p_W = .50$) angelegt, da ab diesem Punkt die Antworten auf mehr Aufgaben gewusst als nicht gewusst werden. In der Medizin findet dieser Ansatz bereits Beachtung. Bei dem dort oft verwendeten *SR*-Aufgabenformat mit fünf Alternativen beträgt die Ratewahrscheinlichkeit $p_R = .20$. Aus der Modellgleichung ergibt sich nun für 50% Wissen ein erwarteter Anteil Punkte an der Maximalpunktzahl von $p_k = .60$, wenn $p_{ftW} = 0$ angenommen wird. Die berechnete Bestehensgrenze liegt also bei 60% der Punkte an der Maximalpunktzahl. Dieser Wert ist in der Medizin bundesweit für das *SR*-Aufgabenformat vorgesehen (s. Bundesministerium der Justiz und für Verbraucherschutz, 2002).

Für die Zusatzaufgaben wurden, auch wenn diese nicht bestehensrelevant waren, dennoch hypothetische Bestehensraten ermittelt, um diese mit den Bestehensraten in den regulären Klausuren zu vergleichen. Dazu wurde die Bestehensgrenze für den jeweiligen Aufgabentyp anhand der obigen Modellgleichung ermittelt, wobei auch hier 50% Wissen als Bestehenskriterium mit $p_{ftW} = 0$ angelegt wurde.

3.2 Prüfung in 2012

3.2.1 Fragestellung und Hypothesen

Im ersten Jahr war die empirische Untersuchung der Zusatzaufgaben stark explorativ angelegt, um zu überprüfen, ob ein *MR-Rating*-Format überhaupt dazu geeignet

ist, in Prüfungen angewendet zu werden. Es stellte sich daher die Frage, inwieweit die Leistung von Prüflingen in den Zusatzaufgaben mit der Leistung in den regulären Klausuren zusammenhängen, welche nur aus *SR*-Aufgaben bestehen. Da sowohl die Aufgaben für die regulären Klausuren als auch die Zusatzaufgaben aus demselben Aufgabenpool entnommen wurden und dasselbe Stoffgebiet abdecken, sollte die Leistung in etwa gleich sein.

3.2.2 Methoden

Versuchsplan

Die unabhängige Variable war das Format, in welchem die einzelnen Prüflinge die Zusatzaufgaben erhielten. Die Hälfte der ausgegebenen Bögen enthielten die Zusatzfragen in ihrer ursprünglichen *SR*-Variante (*SR*-Gruppe) und die andere Hälfte in einer entsprechend den Ausführungen in Absatz 3.1.3 angepassten *MR*-Variante mit einem *Rating*-Antwortformat mit fünf Stufen (*R5*-Gruppe). Die verschiedenen Aufgabenblätter wurden zufällig auf die Prüflinge verteilt.

Als abhängige Variablen wurden die Leistung der Prüflinge in den regulären Prüfungen und die Leistung in den Zusatzaufgaben gemessen.

Stichprobe

An den Prüfungen nahmen insgesamt $N_0 = 245$ Medizinstudierende teil, die sich alle im ersten Jahr des klinischen Medizinstudiums befanden. Alle Prüflinge absolvierten sowohl die Zwischenklausur als auch die Abschlussklausur gemeinsam mit den Zusatzaufgaben. Es konnten allerdings nur die Daten von $N = 188$ Prüflingen, wie weiter unten dargestellt wird, ausgewertet werden. Die restlichen Studierenden mussten von der weiteren Analyse ausgeschlossen werden, da sie entweder das Aufgabenblatt mit

den Zusatzfragen nicht abgaben (1), ihre Matrikelnummer nicht auf dem Aufgabenblatt mit den Zusatzfragen angaben und so ihre Leistung nicht mit der Leistung in den regulären Klausuren verglichen werden konnte (14) oder, in der *R5*-Gruppe, mehr als fünf der insgesamt 50 Alternativen nicht beantworteten (42). Für die *SR*-Gruppe existierte ein ähnliches Kriterium zum Ausschluss von Prüflingen, die nicht genügend Alternativen beantwortet hatten. Dieses musste jedoch nicht angewendet werden, da es keinen Prüfling in der *SR*-Gruppe gab, der eine ganze Aufgabe ausgelassen hatte.

Diese insgesamt $n_{DO} = 57$ von der weiteren Analyse ausgeschlossenen Prüflinge werden im Weiteren als *Drop-Out*-Gruppe (*DO*-Gruppe) bezeichnet. Somit ergaben sich trotz der zufälligen Verteilung der Aufgabenblätter mit $n_{SR} = 129$ für die *SR*-Gruppe und $n_{R5} = 59$ für die *R5*-Gruppe deutlich ungleiche Gruppengrößen. Es gibt jedoch keinen Grund zu der Annahme, dass diesem *Drop-Out* eine besondere Systematik zugrunde lag.

Material und Aufgabe

Bei den Zusatzaufgaben der *SR*-Gruppe handelte es sich um zehn normale Typ-A-Aufgaben, die in früheren Klausuren bereits verwendet wurden. Sie waren in Inhalt und Form mit den Aufgaben in der regulären Abschlussklausur vergleichbar. Zur Beantwortung einer Aufgabe befand sich wie im normalen Typ-A-Format links neben jeder Alternative eine Box, welche bei der richtigen Alternative angekreuzt werden sollte (s. Abbildung 3.3, oben).

Bei den zehn Zusatzaufgaben der *R5*-Gruppe handelte es sich um die gleichen Typ-A-Aufgaben wie in der *SR*-Gruppe, allerdings wurde deren Aufgabenstamm entsprechend der Ausführungen in Abschnitt 3.1.3 angepasst. Bei neun der zehn Fragen wurden die ursprünglichen Alternativen beibehalten, so dass weiterhin genau eine Alter-

SR	Welches der folgenden Antimykotika sollte am ehesten zur Therapie systemischer Pilzinfektionen verwendet?
☒	Amphotericin B
☐	Flucytosin
☐	Ketoconazol
☐	Fluconazol
☐	Griseofulvin

R5	Welche der folgenden Antimykotika können zur Therapie systemischer Pilzinfektionen verwendet werden?					
		sicher richtig	eher richtig	weiß ich nicht	eher falsch	sicher falsch
	Amphotericin B	☒	☐	☐	☐	☐
	Nystatin	☒	☐	☐	☐	☐
	Tolnaftat	☒	☐	☐	☐	☐
	Amorolfin	☒	☐	☐	☐	☐
	Griseofulvin	☐	☐	☐	☐	☒

Abbildung 3.3. Beispielaufgabe mit Antwortschlüssel aus der Prüfung im Jahr 2012 im *SR*-Format (oben) und im *R5*-Format (unten). Die abgedruckten Kreuze dienen zur Information des Lesers und geben die bestmögliche Lösung wieder; alle Antwortboxen auf dem Aufgabenblatt waren selbstverständlich leer.

native korrekt war. Bei der verbleibenden Aufgabe, welche als erste Zusatzaufgabe auf dem Aufgabenblatt erschien, wurde die alte *SR*-Variante negiert, so dass nun die vier ursprünglichen Distraktoren die vier richtigen Alternativen darstellten und die ursprünglich richtige Alternative falsch war. Zur Beantwortung einer Aufgabe befanden sich rechts neben jeder Alternative fünf Boxen, die von links nach rechts mit „sicher richtig“, „eher richtig“, „weiß ich nicht“, „eher falsch“ und „sicher falsch“ betitelt waren (s. Abbildung 3.3, unten). Aufgabe der Prüflinge war es, für jede Alternative, also in jeder Zeile mit fünf Boxen, diejenige Box anzukreuzen, die ihrer Meinung nach am ehesten zutreffend ist.

3.2.3 Ergebnisse

Stichproben

Zwischen den drei Gruppen *SC*, *R5* und *DO* wurden keine Unterschiede hinsichtlich Alter zum Prüfungszeitpunkt, Kruskal-Wallis[6] $\chi^2_2 = .909$, $p = .635$ oder Geschlechtsverteilung, $\chi^2_2 = 1.82$, $p = .401$, festgestellt. Deskriptive Daten hierzu finden sich in Tabelle 3.1. Weiterhin konnte kein Unterschied beim Gesamtpunktwert in den regulären Klausuren, $F(2,242) = .213$, $p = .808$, gefunden werden. Daher können die drei Gruppen als gleichwertig in ihren zugrundeliegenden Eigenschaften betrachtet werden. Das mittels der beiden Klausuren geprüfte Modul wurde von 95.9% der Prüflinge bestanden.

Bewertung mit Punkten

Die Bewertung der Zusatzaufgaben mit Punkten wurde in den beiden Gruppen unterschiedlich gehandhabt: In der *SR*-Gruppe wurde genau wie in den beiden regulären Klausuren ein klassischer Summenwert gebildet. Alle Aufgaben, bei denen die Box der richtigen Alternative korrekterweise angekreuzt war und die Boxen aller anderen Alternativen freigelassen waren, wurden mit einem Punkt bewertet. In allen anderen möglichen Fällen wurde die Aufgabe als nicht korrekt beantwortet angesehen und mit Null Punkten bewertet. Die Punkte der einzelnen Aufgaben wurden gleichgewichtet aufsummiert, so dass eine Maximalpunktzahl von zehn Punkten erreichbar war. Obwohl die Zusatzaufgaben nicht bestehensrelevant für das Modul waren, wur-

[6] Da die Spannweite des Alters der getesteten Stichprobe aufgrund der Beschränkung auf den Universitätskontext stark eingeschränkt ist, kann dieses nicht als normalverteilt angenommen werden. Daher wurde hier und in 2013 auf den Kruskal-Wallis-Test für mehr als zwei unabhängige Stichproben und in 2014 auf den Mann-Whitney-*U*-Test für zwei unabhängige Stichproben als nichtparametrische Verfahren zum Test auf Unterschiede in der zentralen Tendenz zurückgegriffen (Hollander & Wolfe, 1973).

Tabelle 3.1. Deskriptive Daten der Stichprobe für die Untersuchung im Jahr 2012.

	Geschlechtsverteilung		
Gruppe	weiblich	männlich	Alter
SC	82	47	24.6
R5	32	27	24.5
DO	37	20	25.0
Gesamt	151	94	24.7

de dennoch eine hypothetische Bestehensrate ermittelt. Hierzu wurde anhand der in Abschnitt 3.1.4 geschilderten Überlegungen die hypothetische Bestehensgrenze auf 60% der Maximalpunktzahl festgelegt, da die Ratewahrscheinlichkeit in der *SR*-Gruppe $p_R = 1/5 = .20$ betrug.

In der *R5*-Gruppe wurde jede der 50 Alternativen einzeln nach folgendem Schema bewertet: Eine Alternative wurde genau dann als korrekt beantwortet angesehen und mit einem Punkt bewertet, wenn entweder die Alternative richtig war und mit „sicher richtig“ oder „eher richtig“ beantwortet wurde oder wenn die Alternative falsch war und mit „sicher falsch“ oder „eher falsch“ beantwortet wurde. In allen anderen möglichen Fällen wurde die Alternative als nicht korrekt beantwortet angesehen und mit Null Punkten bewertet. Die Punkte der einzelnen Alternativen wurden gleichgewichtet aufsummiert, so dass eine Maximalpunktzahl von 50 Punkten erreicht werden konnte. Auch hier wurde eine hypothetische Bestehensrate ermittelt. Da nun bei jeder einzelnen Alternative jeweils fünf Kategorien zur Beantwortung zur Verfügung standen, von denen jeweils zwei zu einer korrekten Antwort führten, erhöhte sich die Ratewahrscheinlichkeit auf $p_R = 2/5 = .40$. Daher betrug die hypothetische Bestehensgrenze in der *R5*-Gruppe anhand der in Abschnitt 3.1.4 geschilderten Überlegungen nun 70% der Maximalpunktzahl.

Ergebnisse der Punktebewertung

Während 95.6% der Prüflinge das gesamte Modul bestanden, war die Bestehensrate der regulären Abschlussklausur zwar um neun Prozent geringer, lag mit 86.9% jedoch weiterhin recht hoch. Demgegenüber hätten nur 78.3% der Prüflinge in der *SR*-Gruppe die Zusatzaufgaben bestanden, obwohl diese exakt den gleichen Lernstoff mit dem gleichen Aufgabenformat wie die reguläre Abschlussklausur prüften. In der *R5*-Gruppe sind es sogar nur noch 50.8% der Prüflinge, die diese Aufgaben bestanden hätten.

Um den Zusammenhang der Leistungen in den Zusatzaufgaben und der Leistungen in den regulären Klausuren zu ermitteln, wurden Korrelationen bestimmt. Die Korrelation zwischen dem Gesamtpunktwert (GP) aus Zwischen- und Abschlussklausur und dem Punktwert in den Zusatzaufgaben war in der *R5*-Gruppe am größten mit $r = .365$, $p = .004$, und wurde dicht gefolgt von der Korrelation in der *SC*-Gruppe mit $r = .359$, $p < .001$. Es konnte demnach kein signifikanter Unterschied der beiden Korrelationen[7] festgestellt werden, $z = .043$, $p = .483$.

Eine Zusammenfassung der Ergebnisse findet sich in Tabelle 3.2. In Abbildung 3.4 sind Streudiagramme der erreichten Gesamtpunkte in den beiden Klausuren im Verhältnis zu den erreichten Punkten in den Zusatzaufgaben für beide Gruppen dargestellt.

Ergebnisse der Signalentdeckungsparameter

Eine Schätzung der Signalentdeckungsparameter ist in der *SR*-Gruppe nicht sinnvoll, da hier jeweils bei einer Aufgabe aus fünf Alternativen nur eine einzige ausgewählt

[7] Alle in dieser Arbeit berichteten Signifikanztests auf Unterschiede zwischen zwei Korrelationen wurden mittels Fischers *r*-zu-*z*-Transformation durchgeführt (s. J. Cohen & Cohen, 1983, S. 54, Gleichung 2.8.5 und Preacher, 2002).

Tabelle 3.2. Ergebnisse für die reguläre Prüfung und die Zusatzaufgaben im Jahr 2012. Dargestellt sind die Ergebnisse für die reguläre Zwischen- (ZK) und Abschlussklausur (AK) sowie deren Summe (Gesamtpunktwert, GP). Weiterhin sind die Ergebnisse der Zusatzaufgaben in beiden Gruppen *SR* und *R5* sowohl für die Bewertung mit Punkten als auch für die Parameterschätzung nach der Signalentdeckungstheorie (*SDT*) abgedruckt.

	N	Min	Max	*M*	*SD*	Bestehens-rate	*r* mit GP	*r* mit AK
				reguläre Klausuren				
ZK	245	.500	.933	.789	.085	.976	.774***	.402***
AK	245	.167	.967	.705	.119	.869	.891***	1
GP	245	.350	.917	.747	.086	.959	1	.891***
				Zusatzaufgaben, Punkte				
SR	129	0	1	.683	.212	.783	.359***	.308***
R5	59	0	.980	.640	.230	.508	.365**	.417***
				Zusatzaufgaben, *SDT*				
R5								
μ_s	59	-.210	5.289	1.514	.880		.243	.307**
σ_s^2	59	.087	16.151	2.177	2.360			
AUC	59	.457	.971	.795	.130		.359*	.457***

Anmerkungen: ***$p < .001$; **$p < .01$; *$p < .05$

werden musste und dies den Prüflingen bekannt war. Rein praktisch ist es weiterhin möglich, die eine richtige Alternative als das Signal und die vier anderen, falschen Alternativen als das Rauschen aufzufassen. Dies ist jedoch wenig sinnvoll, da ein Prüfling nur bei einer Alternative eine durch die Datenlage nachvollziehbare Entscheidung trifft. Erkennt der Prüfling die richtige Alternative, müssen die falschen Alternativen nicht weiter geprüft werden, so dass dort keine Entscheidung getroffen wird, also weder bewusst ein *false alarm* noch eine *correct rejection* entstehen kann. Es wurden daher nur in der *R5*-Gruppe die Signalentdeckungsparameter mittels des in Kapitel 2 dargestellten *cumulative link model*-Ansatzes in einem *unequal-variance*-Modell geschätzt.

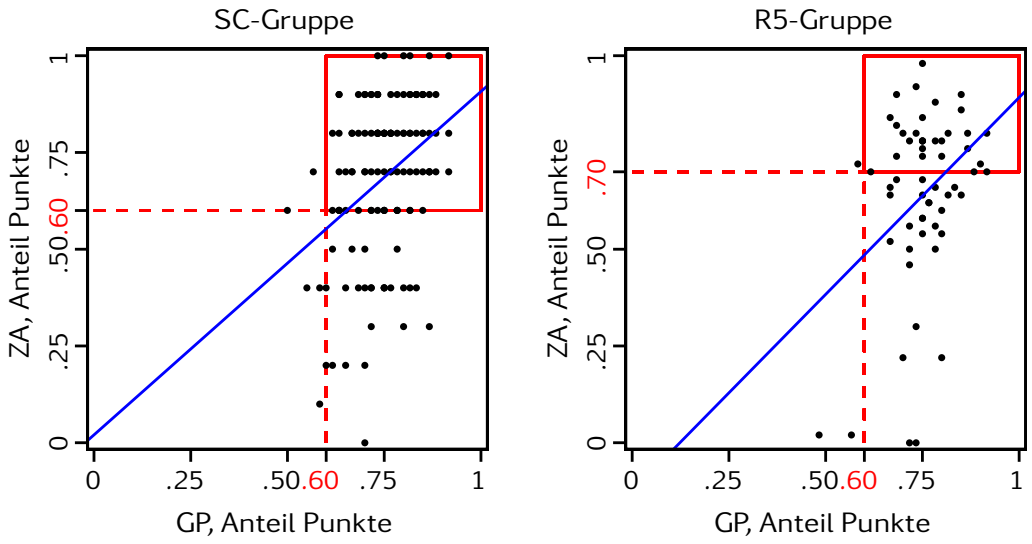

Abbildung 3.4. Gesamtpunkte (GP) in der regulären Prüfung 2012 abgetragen gegen die Punkte in den Zusatzaufgaben (ZA), jeweils als Anteil an der Maximalpunktzahl für die beiden Gruppen *SR* und *R5*. Jeder Punkt im Koordinatensystem stellt die Leistung eines einzelnen Prüflings dar. Rot markiert sind die für die Aufgabenart maßgeblichen Bestehensgrenzen für 50% Wissen. Prüflinge, die sich auf bzw. innerhalb des roten Rechtsecks befinden, haben sowohl die regulären Klausuren als auch die Zusatzaufgaben bestanden.

Wie in Abschnitt 3.2.2 beschrieben, wurden diejenigen Prüflinge von der Auswertung ausgeschlossen, die bei mehr als fünf Alternativen keine Antwort abgegeben haben. Hier ist es jedoch wichtig zu bemerken, dass dies weiterhin dazu führen kann, dass bis zu fünf Antworten für einen Prüfling fehlen. Da das Signalentdeckungsmodell jedoch keine Möglichkeit bietet, fehlenden Datenpunkten Rechnung zu tragen, wurden diese bei der Parameterschätzung verworfen und die Parameter nur an die vorhandenen Daten angepasst, so dass den Schätzungen der einzelnen Prüflinge ggf. leicht unterschiedliche Fallzahlen zugrunde liegen.

Um den Zusammenhang der geschätzten Signalentdeckungsparameter mit den Leistungen in den regulären Klausuren bestimmen zu können, wurden wiederum Korrelationen für μ_s und *AUC* berechnet. Die Korrelation zwischen dem Gesamtpunktwert aus den Klausuren und μ_s ist mit $r = .243$ die geringste aller berechneten Korrelatio-

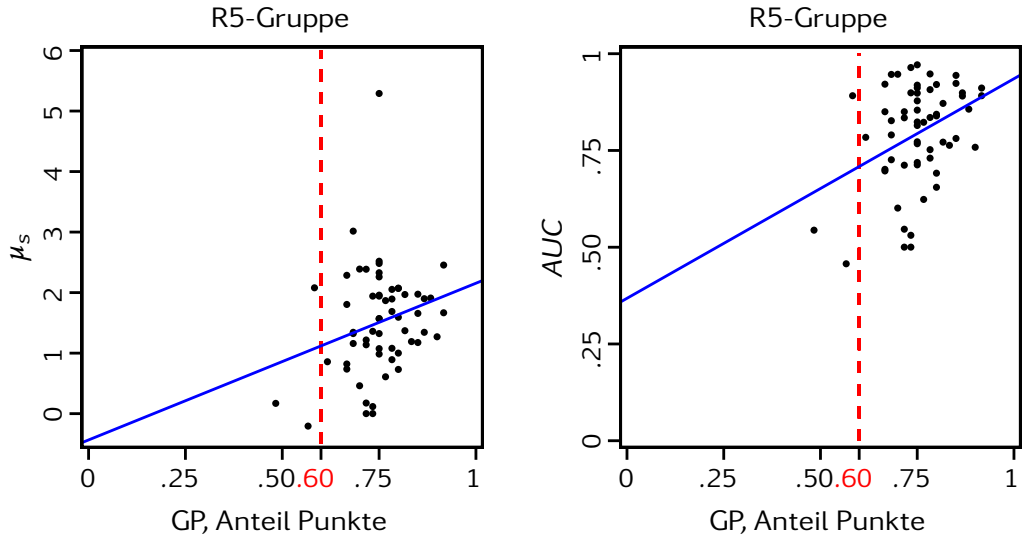

Abbildung 3.5. Anteil der Gesamtpunkte (GP) in der regulären Prüfung 2012 abgetragen gegen die geschätzten Parameter für das Signalentdeckungsmodell in den Zusatzaufgaben für die *R5*-Gruppe. Jeder Punkt im Koordinatensystem stellt die Leistung eines einzelnen Prüflings dar. Rot markiert ist die Bestehensgrenze in der regulären Klausur für 50% Wissen.

nen und nicht signifikant, $p = .064$. Demgegenüber ist die Korrelation der *AUC* mit dem Gesamtpunktwert mit $r = .359$, $p = .005$, signifikant und bis zur dritten Kommastelle genauso groß wie die Korrelation mit dem Gesamtpunktwert bei der Auswertung nach Punkten in den beiden Gruppen, so dass sich ein Signifikanztest an dieser Stelle erübrigt.

Die Ergebnisse für μ_s, σ_s^2 und *AUC* sind auch in Tabelle 3.2 dargestellt. In Abbildung 3.5 sind Streudiagramme der erreichten Gesamtpunkte in den Klausuren im Verhältnis zu den geschätzten Signalentdeckungsparametern in den Zusatzaufgaben dargestellt.

3.2.4 Diskussion

Ziel der Untersuchung war es, die Tauglichkeit eines *Rating*-Verfahrens als Antwortschlüssel für Aufgaben in Prüfungen zu überprüfen und mit dem Standardverfahren

in der Medizin in dieser Hinsicht, dem *SR*-Verfahren mit fünf Alternativen, zu vergleichen.

Bewertung der Ergebnisse

Ein auffälliges Ergebnis ist, dass von denjenigen Prüflingen, welche das *Rating*-Verfahren als Antwortschlüssel in den Zusatzaufgaben erhielten, sehr viel weniger diese Zusatzaufgaben hypothetisch bestanden hätten, als von denjenigen, welche die *SR*-Variante als Antwortschlüssel erhielten (50.8% vs. 78.3%). Dies ist erstaunlich, da es sich jeweils um die gleichen Aufgaben handelte, welche gleichermaßen aus demselben Prüfungsstoff entstammten, der in der regulären Abschlussklausur geprüft wurde. Noch erstaunlicher ist, dass die Korrelationen der erreichten Punkte in den Zusatzaufgaben mit dem Gesamtpunktwert der regulären Prüfungen dennoch für beide Gruppen in etwa gleich groß sind. Erfreulich in dieser Hinsicht ist, dass sich auch die Korrelation der *AUC* mit dem Gesamtpunktwert in ähnlicher Höhe befindet, was einen Hinweis auf deren Anwendbarkeit als Leistungsmaß zur Bewertung von Prüfungen liefert.

Kritisch bleibt jedoch festzustellen, dass selbst in der *SR*-Gruppe, in der ein für die Prüflinge sehr bekanntes und vertrautes Antwortformat benutzt wurde, die Bestehensrate von über 95% auf nicht einmal 80% einbrach. Es lässt sich vermuten, dass die Prüflinge nach der regulären Klausur möglicherweise wenig motiviert waren, die Zusatzaufgaben gewissenhaft zu beantworten, da dies freiwillig und ohne Gegenleistung geschah. Dieser motivationale Aspekt mag auch in der *R5*-Gruppe zu der recht hohen *Drop-Out*-Rate von fast 50% geführt haben, da es sich bei diesem Antwortschlüssel für die Prüflinge um ein völlig neues, unbekanntes Format handelte.

Dennoch bewegten sich die Ergebnisse in die erwartete Richtung, so dass die Idee

eines *Rating*-Verfahrens mit Auswertung nach der Signalentdeckungstheorie in den folgenden Jahren weiter verfolgt wurde.

Methodische Mängel

Bei der kritischen Betrachtung des Bewertungsvorgangs der einzelnen Alternativen mit Punkten beim *Rating*-Verfahren stellte sich dieser als problematisch heraus. Speziell ist hierbei die mittlere Kategorie der Bewertungsskala zu nennen, welche mit „Weiß ich nicht“ betitelt war. Ein Kreuz in dieser Kategorie wird bei jeder Alternative, egal ob tatsächlich richtig oder tatsächlich falsch, mit nicht korrekt bewertet, so dass es sich für einen Prüfling höchst ungünstig auswirkt, hier Kreuze zu setzen. Darüber hinaus ergibt sich ein syntaktisches Problem: Während die Kategorie „Weiß ich nicht“ eine Aussage über den eigenen kognitiven Status eines Prüflings ist, stellen die anderen vier Kategorien „sicher/eher richtig“ bzw. „sicher/eher falsch“ Aussagen über die einzelnen Alternativen dar und sind somit nicht auf der gleichen kognitiven Dimension lokalisierbar. Daher sollte die Anzahl der Kategorien für ein *Rating*-Verfahren als Antwortschlüssel in Prüfungen geradzahlig sein. Dies gilt selbst dann, wenn die mittlere Kategorie ohne Titel bleibt, da einerseits weiterhin das Problem der Bewertung verbleibt und andererseits, um auszuschließen, dass Prüflinge diese Kategorie von selbst immer dann wählen, wenn sie die Antwort nicht kennen.

Weiterhin stellte sich im Nachhinein heraus, dass die Auswahl der Aufgaben für die Zusatzfragen ungünstig war, da einerseits Abhängigkeiten zwischen zwei der Aufgaben bestanden und andererseits teilweise Hinweise innerhalb einer Aufgabe das Lösen einer anderen Aufgabe ohne Wissen begünstigten (vgl. Aufgabe 3 und 4 in Anhang B.1 und B.2). Daher wurden für die Untersuchung im Jahr 2013 neue Aufgaben ausgewählt.

3.3 Prüfung in 2013

3.3.1 Fragestellung

Im zweiten Jahr der empirischen Untersuchung lag der Fokus auf dem Vergleich der drei in Kapitel 1 angesprochenen Antwortschlüssel für das *MR*-Aufgabenformat (vgl. Abbildung 1.1): *multiple choice* (*MC*), *multiple true-false* (*MTF*) und erneut einem *Rating* mit vierstufiger Antwortskala (*R4*). Da es zwischen den Antwortschlüsseln keine formalen Unterschiede, z.B. hinsichtlich der Ratewahrscheinlichkeit gibt und auch die Vertrautheit der Prüflinge mit allen drei Varianten als gering eingeschätzt werden kann, sollten sich keine Unterschiede hinsichtlich der Leistungen ergeben.

Weiterhin sollte eine Einschätzung durch die Prüflinge bezüglich der Eignung des *MR*-Aufgabenformats und der verschiedenen Antwortschlüssel für Prüfungen und in der Lehre gewonnen werden. Es wurde erwartet, dass das neue Aufgabenformat als ungewohnt und daher schwierig eingeschätzt wird, es jedoch dahingehend keine Unterschiede zwischen den einzelnen Antwortschlüsseln geben sollte.

3.3.2 Methoden

Versuchsplan

Die unabhängige Variable war das Format, in welchem die einzelnen Prüflinge die Zusatzaufgaben erhielten. Alle ausgegebenen Bögen enthielten die Zusatzfragen im, entsprechend den Ausführungen in Absatz 3.1.3 angepassten, *MR*-Format, davon jeweils ein Drittel mit *MC*-, *MTF*- bzw. *R4*-Antwortschlüssel. Die verschiedenen Aufgabenblätter wurden zufällig auf die Prüflinge verteilt, woraus sich jeweils deren Gruppenzugehörigkeit ergab.

Als abhängige Variablen wurden die Leistung der Prüflinge in den regulären Prüfungen und die Leistung in den Zusatzaufgaben gemessen.

Stichprobe

An der Zwischenklausur nahmen insgesamt $N_{0,ZK} = 218$ und an der Abschlussklausur insgesamt $N_{0,AK} = 211$ Medizinstudierende teil, die sich alle im ersten Jahr des klinischen Medizinstudiums befanden. Prüflinge, die nicht an beiden Klausuren teilnahmen, wurden von der weiteren Auswertung ausgeschlossen. Es verblieben $N_0 = 209$ Prüflinge, die sowohl die Zwischenklausur als auch die Abschlussklausur gemeinsam mit den Zusatzaufgaben absolvierten.

Es konnten allerdings nur die Daten von $N = 162$ Prüflingen, wie weiter unten dargestellt wird, ausgewertet werden. Die restlichen Prüflinge mussten von der weiteren Analyse ausgeschlossen werden, da sie entweder das Aufgabenblatt mit den Zusatzfragen nicht abgaben (5), ihre Matrikelnummer nicht auf dem Aufgabenblatt mit den Zusatzfragen angaben und so ihre Leistung nicht mit der Leistung in den regulären Klausuren verglichen werden konnte (2), in der *MC*-Gruppe eine ganze Aufgabe nicht beantworteten (5) oder in der *MTF*-Gruppe (14) bzw. in der *R4*-Gruppe (21) mehr als fünf der insgesamt 50 Alternativen nicht beantworteten.

Diese insgesamt $n_{DO} = 47$ von der weiteren Analyse ausgeschlossenen Prüflinge werden im Weiteren als *Drop-Out*-Gruppe (*DO*-Gruppe) bezeichnet, wobei es keinen Grund zu der Annahme gibt, dass diesem *Drop-Out* eine besondere Systematik zugrunde lag. Somit ergaben sich für die drei anderen Gruppen mehr oder weniger gleich große Gruppengrößen von $n_{MC} = 58$ für die *MC*-Gruppe, $n_{MTF} = 56$ für die *MTF*-Gruppe und $n_{R4} = 48$ für die *R4*-Gruppe.

Material und Aufgabe

Bei den Zusatzaufgaben handelte es sich um zehn Typ-A-Aufgaben, die in früheren Klausuren bereits verwendet wurden. Sie waren in ihrem Inhalt mit den Aufgaben in den regulären Klausuren vergleichbar, allerdings wurde deren Aufgabenstamm entsprechend der Ausführungen in Abschnitt 3.1.3 angepasst.

Weiterhin wurden die Alternativen so gestaltet, dass nun bei jeweils zwei Aufgaben eine, zwei, drei, vier oder alle fünf Alternativen richtig waren. Von der Verwendung von Aufgaben, bei denen keine Alternative richtig war, wurde abgesehen, da dies einerseits technisch in EvaExam (Electric Paper Evaluationssysteme GmbH, 2015a) bei *MC*-Aufgaben nicht möglich ist und andererseits Prüflinge möglicherweise stark verunsichert sein könnten, wenn sie in einer ganzen *MC*-Aufgabe kein einziges Kreuz setzen müssen.

Zur Beantwortung einer Aufgabe in der *MC*-Gruppe befand sich links neben jeder Alternative eine Box, welche bei einer richtige Alternative angekreuzt und bei einer falschen Alternative freigelassen werden sollte (s. Abbildung 3.6, oben).

Zur Beantwortung einer Aufgabe in der *MTF*-Gruppe befanden sich rechts neben jeder Alternative zwei Boxen, wovon die von linke Box mit „richtig“ und die rechte Box mit „falsch“ betitelt waren (s. Abbildung 3.6, Mitte). Aufgabe der Prüflinge war es, für jede Alternative zu entscheiden, ob diese richtig oder falsch ist und in der zugehörigen Zeile mit zwei Boxen die entsprechende Antwort anzukreuzen.

Zur Beantwortung einer Aufgabe in der *R4*-Gruppe befanden sich rechts neben jeder Alternative vier Boxen, die von links nach rechts mit „sicher richtig“, „eher richtig“, „eher falsch“ und „sicher falsch“ betitelt waren (s. Abbildung 3.6, unten). Aufgabe der Prüflinge war es, für jede Alternative, also in jeder Zeile mit vier Boxen, diejeni-

MC	Welche der folgenden Substanzen ist sinnvoll für die Behandlung der Myasthenia gravis?
	☐ Diazepam
	☐ Atropin
	☒ Neostigmin
	☐ Scopolamin
	☒ Pyridostigmin

MTF	Welche der folgenden Substanzen ist sinnvoll für die Behandlung der Myasthenia gravis?	
	richtig	falsch
Diazepam	☐	☒
Atropin	☐	☒
Neostigmin	☒	☐
Scopolamin	☐	☒
Pyridostigmin	☒	☐

R4	Welche der folgenden Substanzen ist sinnvoll für die Behandlung der Myasthenia gravis?			
	sicher richtig	eher richtig	eher falsch	sicher falsch
Diazepam	☐	☐	☐	☒
Atropin	☐	☐	☐	☒
Neostigmin	☒	☐	☐	☐
Scopolamin	☐	☐	☐	☒
Pyridostigmin	☒	☐	☐	☐

Abbildung 3.6. Beispielaufgabe mit Antwortschlüssel aus der Prüfung im Jahr 2013 im *MC*-Format (oben), im *MTF*-Format (Mitte) und im *R4*-Format (unten). Die abgedruckten Kreuze dienen zur Information des Lesers und geben die bestmögliche Lösung wieder; alle Antwortboxen auf dem Aufgabenblatt waren selbstverständlich leer.

ge Box anzukreuzen, die ihrer Meinung nach am ehesten zutreffend ist.

3.3.3 Ergebnisse

Stichproben

Zwischen den vier Gruppen *MC*, *MTF*, *R4* und *DO* wurden keine Unterschiede hinsichtlich Alter zum Prüfungszeitpunkt, Kruskal-Wallis $\chi^2_3 = .150$, $p = .985$ oder Geschlechtsverteilung, $\chi^2_3 = 1.54$, $p = .673$, festgestellt. Deskriptive Daten hierzu finden sich in Tabelle 3.3. Weiterhin konnte kein Unterschied beim Gesamtpunktwert in den regulären Klausuren, $F(3, 205) = 1.61$, $p = .188$, gefunden werden. Daher können die vier Gruppen als gleichwertig in ihren zugrundeliegenden Eigenschaften betrachtet werden.

Für alle weiteren Betrachtungen ist es notwendig, darauf hinzuweisen, dass das mittels der beiden Klausuren geprüfte Modul nur von 69.4% der Prüflinge bestanden wurde. Dies sind deutlich weniger als noch im Jahr 2012, als 95.9% der Prüflinge das Modul bestanden. Während die Zwischenklausur noch von 78.0% der Prüflinge bestanden wurde (2012: 97.6%), war die Bestehensrate in der Abschlussklausur, in deren Anschluss die Zusatzaufgaben bearbeitet wurden, auf nur noch 48.8% abgesunken (2012: 86.9%). Es ist daher davon auszugehen, dass dieser Sachverhalt die nachfolgenden Ergebnisse stark negativ beeinflusst haben könnte.

Bewertung mit Punkten

Die Bewertung der Zusatzaufgaben mit Punkten wurde in den drei Gruppen unterschiedlich gehandhabt. Da es sich im Jahr 2013 allerdings in allen Gruppen um Aufgaben im *MR*-Format handelte, wurde, wie bereits bei der Untersuchung im Vorjahr bei diesem Format, jede der 50 Alternativen einzeln bewertet.

In der *MC*-Gruppe wurde eine Alternative genau dann als korrekt beantwortet an-

Tabelle 3.3. Deskriptive Daten der Stichprobe für die Untersuchung im Jahr 2013.

	Geschlechtsverteilung		
Gruppe	weiblich	männlich	Alter
MC	40	18	24.3
MTF	36	20	24.2
R4	31	17	23.9
DO	42	16	24.1
Gesamt	149	71	24.1

gesehen und mit einem Punkt bewertet, wenn entweder die Alternative richtig war und die Antwortbox vor der Alternative „angekreuzt“ war oder wenn die Alternative falsch war und die Antwortbox „freigelassen“ wurde. In allen anderen möglichen Fällen wurde die Alternative als nicht korrekt beantwortet angesehen und mit Null Punkten bewertet.

In der *MTF*-Gruppe wurde eine Alternative genau dann als korrekt beantwortet angesehen und mit einem Punkt bewertet, wenn entweder die Alternative richtig war und mit „richtig“ beantwortet wurde oder wenn die Alternative falsch war und mit „falsch“ beantwortet wurde. In allen anderen möglichen Fällen wurde die Alternative als nicht korrekt beantwortet angesehen und mit Null Punkten bewertet.

In der *R4*-Gruppe wurde eine Alternative genau dann als korrekt beantwortet angesehen und mit einem Punkt bewertet, wenn entweder die Alternative richtig war und mit „sicher richtig“ oder „eher richtig“ beantwortet wurde oder wenn die Alternative falsch war und mit „sicher falsch“ oder „eher falsch“ beantwortet wurde. In allen anderen möglichen Fällen wurde die Alternative als nicht korrekt beantwortet angesehen und mit Null Punkten bewertet.

Die Punkte der einzelnen Alternativen wurden gleichgewichtet aufsummiert, so dass

eine Maximalpunktzahl von 50 Punkten erreicht werden konnte. Auch hier wurde für alle Gruppen eine hypothetische Bestehensrate ermittelt. In der *MC*-Gruppe standen zwei Möglichkeiten zur Beantwortung einer Alternative zur Verfügung: „Kreuz“ oder „kein Kreuz“; in der *MTF*-Gruppe standen auch zwei Möglichkeiten zur Beantwortung einer Alternative zur Verfügung: „richtig“ oder „falsch“; in der *R4*-Gruppe standen zur Beantwortung einer Alternative vier Kategorien zur Verfügung, von denen jeweils zwei zu einer korrekten Antwort führten. Es ergibt sich daher für alle drei Gruppen eine Ratewahrscheinlichkeit von $p_R = 1/2 = .50$. Daher betrug die hypothetische Bestehensgrenze in allen drei Gruppen anhand der in Abschnitt 3.1.4 geschilderten Überlegungen nun 75% der Maximalpunktzahl.

Ergebnisse der Punktebewertung

Während 69.4% der Prüflinge das gesamte Modul bestanden haben, sank bereits die Bestehensrate der regulären Abschlussklausur deutlich um mehr als 20% ab, so dass mit 48.8% nicht einmal die Hälfte der Prüflinge diese reguläre Klausur bestanden haben. Dies wirkte sich sehr stark auf die Ergebnisse in den freiwillig zu beantwortenden Zusatzaufgaben aus: Hier hätten nur 5.2% der Prüflinge in der *MC*-Gruppe, 10.7% in der *MTF*-Gruppe und 10.7% in der *R4*-Gruppe die Zusatzaufgaben bestanden.

Um den Zusammenhang der Leistungen in den Zusatzaufgaben und der Leistungen in den regulären Klausuren zu ermitteln, wurden Korrelationen bestimmt. Die Korrelation zwischen dem Gesamtpunktwert (GP) aus Zwischen- und Abschlussklausur und dem Punktwert in den Zusatzaufgaben, war in der *MC*-Gruppe am größten mit $r = .343$ und signifikant, $p = .004$. Die Korrelationen in den anderen beiden Gruppen lagen nahe Null und sind dementsprechend nicht signifikant: $p = .673$ für die *MTF*-Gruppe und $p = .574$ für die *R4*-Gruppe. Sie unterscheiden sich jedoch signifikant von der Korre-

lation in der *MC*-Gruppe mit $z = 2.175$, $p = .015$ (*MTF*), und $z = 1.918$, $p = .028$ (*R4*). Eine Zusammenfassung der Ergebnisse findet sich in Tabelle 3.4. In Abbildung 3.7 sind Streudiagramme der erreichten Gesamtpunkte in den beiden Klausuren im Verhältnis zu den erreichten Punkten in den Zusatzaufgaben für beide Gruppen dargestellt.

Ergebnisse der Signalentdeckungsparameter

Wie in Abschnitt 3.3.2 beschrieben, wurden diejenigen Prüflinge von der Auswertung ausgeschlossen, die bei mehr als fünf Alternativen keine Antwort abgegeben haben. Hier ist es jedoch wichtig zu bemerken, dass dies weiterhin dazu führen kann, dass bis zu fünf Antworten für einen Prüfling fehlen. Da das Signalentdeckungsmodell jedoch keine Möglichkeit bietet, fehlenden Datenpunkten Rechnung zu tragen, wurden diese bei der Parameterschätzung verworfen und die Parameter nur an die vorhandenen Daten angepasst, so dass den Schätzungen der einzelnen Prüflinge ggf. leicht unterschiedliche Fallzahlen zugrunde liegen.

In den beiden Gruppen *MC* und *MTF* stehen den Prüflingen nur zwei Antwortkategorien zur Verfügung, um eine Alternative zu beantworten. Demnach ist es hier nicht möglich, eine Schätzung für die drei Parameter μ_s, σ_s^2 und λ vorzunehmen (vgl. Abschnitt 2.3). Für diese beiden Gruppen wurde daher ein *equal-variance*-Modell mit entsprechender Notation an die Daten angepasst. In der *R4*-Gruppe wurde dagegen wie im Vorjahr in der *R5*-Gruppe ein *unequal-variance*-Modell angepasst.

Um den Zusammenhang der Signalentdeckungsmaße mit den Leistungen in den regulären Klausuren bestimmen zu können, wurden wiederum Korrelationen für μ_s und *AUC* berechnet. In der *MC*-Gruppe ist die Korrelation zwischen dem Gesamtpunktwert aus den Klausuren und d' mit $r = .402$ die größte aller berechneten Korrelationen und signifikant, $p = .002$. Sie wird dicht gefolgt von der Korrelation des Gesamtpunktwerts

Tabelle 3.4. Ergebnisse für die reguläre Prüfung und die Zusatzaufgaben im Jahr 2013. Dargestellt sind die Ergebnisse für die reguläre Zwischen- (ZK) und Abschlussklausur (AK) sowie deren Summe (Gesamtpunktwert, GP). Weiterhin sind die Ergebnisse der Zusatzaufgaben in den drei Gruppen *MC, MTF* und *R4* sowohl für die Bewertung mit Punkten als auch für die Parameterschätzung nach der Signalentdeckungstheorie (*SDT*) abgedruckt.

	N	Min	Max	*M*	*SD*	Bestehens-rate	*r* mit GP	*r* mit AK
				reguläre Klausuren				
ZK	209	.333	.967	.686	.126	.780	.837***	.404***
AK	209	.167	.833	.572	.126	.488	.839***	1
GP	209	.283	.833	.629	.106	.694	1	.839***
				Zusatzaufgaben, Punkte				
MC	58	.440	.780	.585	.081	.052	.343**	.316**
MTF	56	.380	.880	.616	.101	.107	-.061	-.054
R4	48	.460	.800	.637	.080	.104	-.028	.138
				Zusatzaufgaben, *SDT*				
MC								
d'	58	-.016	2.062	.855	.454		.402**	.348**
AUC	58	.496	.928	.717	.100		.390*	.328*
MTF								
d'	56	-.661	2.237	.687	.601		.106	.146
AUC	56	.333	.943	.671	.133		.093	.138
R4								
μ_s	48	-.181	1.573	.754	.394		.124	.177
σ_s^2	48	.004	6.039	1.051	.964			
AUC	48	.454	.888	.700	.096		.136	.163

Anmerkungen: ***$p < .001$; **$p < .01$; *$p < .05$

aus den Klausuren und der *AUC* als der gleichen Gruppe mit $r = .390$, welche ebenfalls signifikant ist $p = .002$.

Die gleichen zwei Korrelationen wurden auch für die anderen beiden Gruppen bestimmt. Diese liegen in der *MTF*-Gruppe bei $r = .106$, $p = .436$ für d' und $r = .093$,

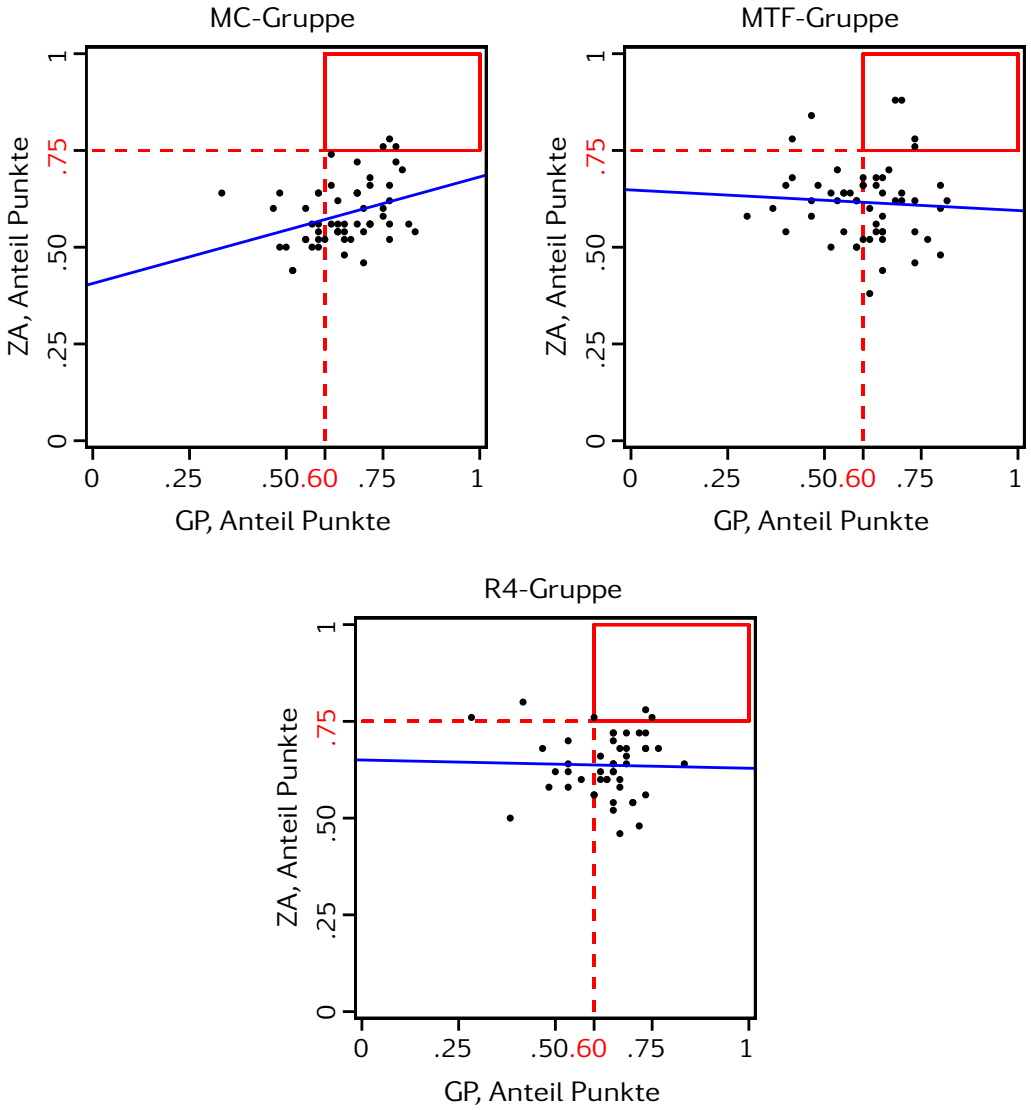

Abbildung 3.7. Gesamtpunkte (GP) in der regulären Prüfung 2013 abgetragen gegen die Punkte in den Zusatzaufgaben (ZA), jeweils als Anteil an der Maximalpunktzahl für die drei Gruppen *MC*, *MTF* und *R4*. Jeder Punkt im Koordinatensystem stellt die Leistung eines einzelnen Prüflings dar. Rot markiert sind die für die Aufgabenart maßgeblichen Bestehensgrenzen für 50% Wissen. Prüflinge, die sich auf bzw. innerhalb des roten Rechtsecks befinden, haben sowohl die regulären Klausuren als auch die Zusatzaufgaben bestanden.

$p = .495$ für die *AUC* und sind demnach nicht signifikant. Ein ähnliches Bild ergibt sich für die Korrelationen in der *R4*-Gruppe mit $r = .124$, $p = .402$ für μ_s und $r = .136$, $p = .356$ für die *AUC*, welche ebenfalls nicht signifikant sind.

Daraufhin wurde auf Unterschiede zwischen den gerade berichteten Korrelationen getestet. Es zeigte sich ein signifikanter Unterschied zwischen den Korrelationen der *MC*-Gruppe und der *MTF*-Gruppe sowohl hinsichtlich der Korrelation von μ_s mit $z = 1.66$, $p = .048$ als auch hinsichtlich der Korrelation der *AUC* mit $z = 1.66$, $p = .049$. Demgegenüber konnte kein Unterschied der Korrelationen zwischen der *MC*-Gruppe und *R4*-Gruppe gefunden werden: $z = 1.50$, $p = .067$ für μ_s und $z = 1.37$, $p = .086$ für die *AUC*. Auf einen Signifikanztest der Korrelationen der *MTF*- und *R4*-Gruppe wurde aufgrund des geringen numerischen Unterschieds verzichtet.

Die Ergebnisse für d' bzw. μ_s, wenn zutreffend, σ_s^2 und *AUC* sind auch in Tabelle 3.4 dargestellt. In Abbildung 3.8 sind Streudiagramme der erreichten Gesamtpunkte in den Klausuren im Verhältnis zu den geschätzten Signalentdeckungsparametern in den Zusatzaufgaben dargestellt.

3.3.4 Diskussion

Ziel der Untersuchung war es, die drei in Kapitel 1 angesprochenen Antwortschlüssel für das *MR*-Aufgabenformat, *MC*, *MTF* und *R4*, zu vergleichen und Rückschlüsse auf deren Eignung für Prüfungen zu ziehen.

Bewertung der Ergebnisse

Das auffälligste Ergebnis der Untersuchung ist, dass die hypothetischen Bestehensraten in den Zusatzaufgaben auf ein Niveau abfielen, bei dem eine normale reguläre Klausur wahrscheinlich hätte annuliert und wiederholt werden müssen. Dies erschwert die Interpretation der Ergebnisse, wenn es diese nicht gar unmöglich macht.

Dennoch scheint es zumindest in der *MC*-Gruppe einen gewissen Zusammenhang zwischen den Leistungen in der regulären Prüfung und den Zusatzaufgaben zu geben.

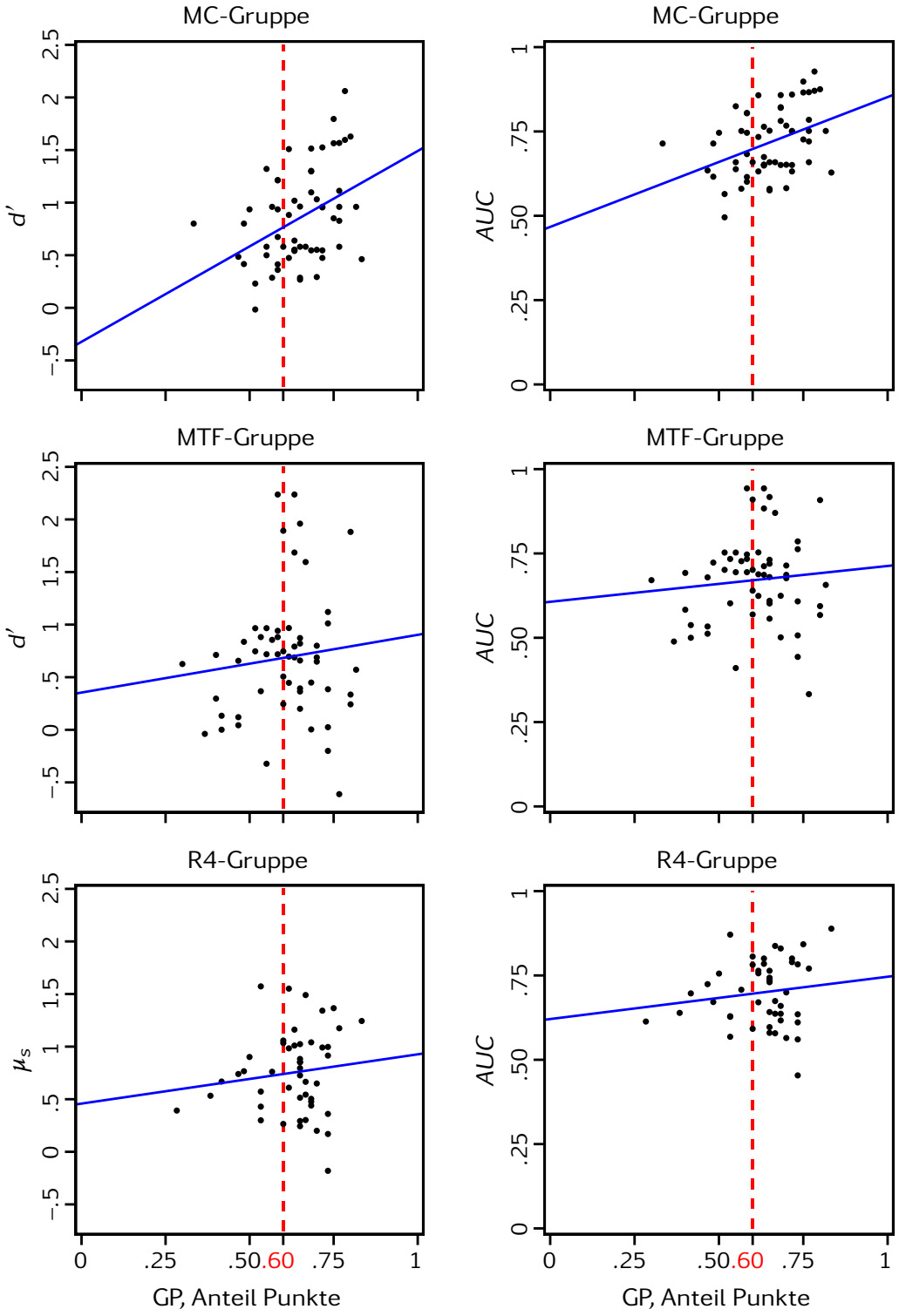

Abbildung 3.8. Anteil der Gesamtpunkte (GP) in der regulären Prüfung 2013 abgetragen gegen die geschätzten Parameter für das Signalentdeckungsmodell in den Zusatzaufgaben für die drei Gruppen *MC, MTF* und *R4*. Jeder Punkt im Koordinatensystem stellt die Leistung eines einzelnen Prüflings dar. Rot markiert ist die Bestehensgrenze in der regulären Klausur für 50% Wissen.

Methodische Mängel

Wie sich bereits im Jahr 2012 angedeutet hat, scheint die Motivation der Prüflinge ein wichtiger Faktor bei der Beantwortung der Zusatzaufgaben zu sein. Da diese eine freiwillige Zusatzleistung ohne direkten Nutzen für die Prüflinge darstellt, könnte man vermuten, dass die Prüflinge wenig motiviert sind, hier ihre tatsächlich mögliche Leistung an den Tag zu legen.

Hinzu kommt, dass das Format der Zusatzaufgaben für die Prüflinge bis zum Prüfungszeitpunkt unbekannt ist. Daher sind die schlechteren Leistungen hier möglicherweise zum Teil auch mit Problemen beim Umgang mit den einzelnen Formaten erklärbar. Insbesondere beim *Rating* in der *R4*-Gruppe könnte Unsicherheit dahingehend bestehen, wie welche Art von Antwort, speziell die beiden „eher“-Kategorien, durch den Prüfer bewertet wird. Dies sollte daher frühzeitig kommuniziert werden, so dass den Prüflingen bereits vor der Prüfung bekannt ist, welche Anforderungen an sie gestellt werden. Weiterhin ist ein gewisser Anreiz für die bestmögliche Bearbeitung der Zusatzaufgaben zu überdenken.

3.4 Prüfung in 2014

3.4.1 Fragestellung

Im dritten Jahr der empirischen Untersuchung sollte der Einfluss der Motivation der Prüflinge auf die Aussagekraft der Ergebnisse untersucht werden, da sich in den beiden vorangegangenen Untersuchungen herausgestellt hatte, dass einerseits ein großer Teil der Prüflinge die Zusatzaufgaben überhaupt nicht bearbeitet hatten und die verbleibenden Prüflinge hier sehr schlechte Leistungen zeigten.

Deshalb wurden die Prüflinge einerseits bereits weit im Voraus über das Aufgaben-

format, die Möglichkeiten zur Beantwortung und die Bewertung der einzelnen Antworten durch den Prüfer informiert. Dies machte es notwendig, nur ein einziges Aufgabenformat für die Zusatzaufgaben zu verwenden, das alle Prüflinge erhalten. Andererseits wurden als Anreiz, die Zusatzaufgaben möglichst gewissenhaft, vollständig und korrekt zu bearbeiten, Bonuspunkte vergeben, welche auf die Leistung in der regulären Klausur angerechnet wurden. Es wurde erwartet, dass diese Maßnahmen einerseits die bisher stets hohe *Drop-Out*-Rate verringern und andererseits die Leistungen der Prüflinge in den Zusatzaufgaben insgesamt verbessern.

3.4.2 Methoden

Versuchsplan

Als unabhängige Variable diente die Zugehörigkeit der einzelnen Prüflinge zur diesjährigen Kohorte 2014 bzw. zur Kohorte im vorherigen Jahr 2013.

Als abhängige Variablen wurden die Leistung der Prüflinge in den regulären Prüfungen und die Leistung in den Zusatzaufgaben gemessen.

Stichprobe

An der Zwischenklausur nahmen insgesamt $N_{0,ZK} = 203$ und an der Abschlussklausur insgesamt $N_{0,AK} = 201$ Medizinstudierende teil, die sich alle im ersten Jahr des klinischen Medizinstudiums befanden. Prüflinge, die nicht an beiden Klausuren teilnahmen, wurden von der weiteren Auswertung ausgeschlossen, so dass $N_0 = 201$ Prüflinge verblieben, die sowohl die Zwischenklausur als auch die Abschlussklausur gemeinsam mit den Zusatzaufgaben absolvierten.

Es konnten allerdings nur die Daten von $N = 194$ Prüflingen, wie weiter unten dargestellt wird, ausgewertet werden. Die restlichen sieben Prüflinge mussten von der wei-

teren Analyse ausgeschlossen werden, da sie mehr als fünf der insgesamt 50 Alternativen nicht beantworteten. Diese $n_{DO} = 7$ von der weiteren Analyse ausgeschlossenen Prüflinge werden im Weiteren als *Drop-Out*-Gruppe (*DO*-Gruppe) bezeichnet, wobei es keinen Grund zu der Annahme gibt, dass diesem *Drop-Out* eine besondere Systematik zugrunde lag.

Material und Aufgabe

Bei den Zusatzaufgaben handelte es sich, bis auf eine Ausnahme (s. Abschnitt B.4), um die gleichen zehn Aufgaben, die bereits im Jahr 2013 benutzt wurden. Dabei handelte es sich um Typ-A-Aufgaben, deren Aufgabenstamm entsprechend der Ausführungen in Abschnitt 3.1.3 angepasst wurde. Weiterhin waren wieder, wie bereits im Vorjahr, bei jeweils zwei Aufgaben eine, zwei, drei, vier oder alle fünf Alternativen richtig.

Zur Beantwortung einer Aufgabe befanden sich rechts neben jeder Alternative vier Boxen, die von links nach rechts mit „sicher richtig“, „eher richtig“, „eher falsch“ und „sicher falsch“ betitelt waren (vgl. Abbildung 3.6, unten). Aufgabe der Prüflinge war es, für jede Alternative, also in jeder Zeile mit vier Boxen, diejenige Box anzukreuzen, die ihrer Meinung nach am ehesten zutreffend ist.

Anders als in den Vorjahren erhielten die Prüflinge den Begleittext zu den Zusatzfragen (vgl. Abschnitt C.3) nicht erst auf dem Aufgabenblatt mit den Zusatzfragen, sondern bereits bei der Prüfungsankündigung ca. dreieinhalb Monate vor der Klausur. Sie wurden in dem Begleittext darüber informiert, in welchem Antwortformat die Zusatzfragen gestaltet sind, wie dieses bearbeitet werden sollte, wie die Bewertung der einzelnen Alternativen durch den Prüfer vorgenommen wird und dass sie durch eine entsprechende Leistung in den Zusatzaufgaben Bonuspunkte erhalten können,

welche auf das Ergebnis der regulären Prüfung angerechnet werden.

3.4.3 Ergebnisse

Stichproben

Zwischen den beiden Gruppen *R4* und *DO* wurden keine Unterschiede hinsichtlich Alter zum Prüfungszeitpunkt, Mann-Whitney-$U = 834$, $z = -1.050$, $p = .294$ oder Geschlechtsverteilung, $\chi^2_1 = 1.02$, $p = .312$, festgestellt. Deskriptive Daten hierzu finden sich in Tabelle 3.5. Weiterhin konnte kein Unterschied beim Gesamtpunktwert in den regulären Klausuren, $t(199) = .663$, $p = .508$, gefunden werden. Daher können die beiden Gruppen als gleichwertig in ihren zugrundeliegenden Eigenschaften betrachtet werden.

Für die weiteren Betrachtungen ist es notwendig, darauf hinzuweisen, dass zwar die Bestehensrate für das mittels der beiden Klausuren geprüfte Modul wieder auf 84.6% anstieg, gegenüber 69.4% im Vorjahr, weiterhin jedoch unter dem Wert von 95.9% im Jahr 2012 lag. Dies ist jedoch einzig auf eine höhere Bestehensrate in der Zwischenklausur von 93.0% zurückzuführen (2012: 97.6%; 2013: 78.0%), während die Bestehensrate in der Abschlussklausur, in deren Anschluss die Zusatzaufgaben bearbeitet wurden, weiter auf nur noch 44.8% absank (2012: 86.9%; 2013: 48.8%). Es ist daher davon auszugehen, dass dieser Sachverhalt die nachfolgenden Ergebnisse stark negativ beeinflusst haben könnte.

Bewertung mit Punkten

Die Bewertung der Zusatzaufgaben mit Punkten wurde für das *R4*-Format wie im Vorjahr gehandhabt. Es wurde wiederum jede der 50 Alternativen einzeln bewertet. Dabei wurde eine Alternative genau dann als korrekt beantwortet angesehen und mit

Tabelle 3.5. Deskriptive Daten der Stichprobe für die Untersuchung im Jahr 2014.

	Geschlechtsverteilung		
Gruppe	weiblich	männlich	Alter
R4	115	79	24.4
DO	8	1	26.0
Gesamt	123	80	24.5

einem Punkt bewertet, wenn entweder die Alternative richtig war und mit „sicher richtig“ oder „eher richtig“ beantwortet wurde oder wenn die Alternative falsch war und mit „sicher falsch“ oder „eher falsch“ beantwortet wurde. In allen anderen möglichen Fällen wurde die Alternative als nicht korrekt beantwortet angesehen und mit Null Punkten bewertet.

Weiterhin wurden die 50 Alternativen zur Bestimmung der Bonuspunkte, die ein Prüfling für die reguläre Klausur erhält, nochmals mit dem sogenannten Bonuspunkt-*Scoring* (BP) bewertet. Dieses wurde den Prüflingen in den Begleittexten zu den Zusatzaufgaben bei der Prüfungsankündigung und auf dem Aufgabenblatt mitgeteilt. Hierbei wurden für eine Alternative zwischen Null und drei Punkten nach dem folgenden Schlüssel vergeben: War eine Alternative richtig, wurden drei Punkte vergeben, wenn mit „sicher richtig“ geantwortet wurde, zwei Punkte, wenn mit „eher richtig“ geantwortet wurde, ein Punkt, wenn mit „eher falsch“ geantwortet wurde und Null Punkte, wenn mit „sicher falsch“ geantwortet wurde. War eine Alternative falsch, wurden drei Punkte vergeben, wenn mit „sicher falsch“ geantwortet wurde, zwei Punkte, wenn mit „eher falsch“ geantwortet wurde, ein Punkt, wenn mit „eher richtig“ geantwortet wurde und Null Punkte, wenn mit „sicher richtig“ geantwortet wurde. Dieses Vorgehen beim Bonuspunkt-*Scoring* führt dazu, dass sich der Erwartungswert für den Anteil der erreichten Punkte an der Maximalpunktzahl im Vergleich zum *R4*-Format nicht ändert.

Die Punkte der einzelnen Alternativen wurden gleichgewichtet aufsummiert, so dass die Maximalpunktzahl für das *R4*-Format weiterhin bei 50 Punkten lag, jenes des Bonuspunkt-*Scorings* bei 150 Punkten. Auch hier wurde für die *R4*-Gruppe eine hypothetische Bestehensrate ermittelt, wobei die Ratewahrscheinlichkeit weiterhin bei $p_R = 1/2 = .50$ lag. Daher betrug die hypothetische Bestehensgrenze in der *R4*-Gruppe anhand der in Abschnitt 3.1.4 geschilderten Überlegungen 75% der Maximalpunktzahl. Diese Bestehensgrenze wurde auch für das Bonuspunkt-*Scoring* zugrundegelegt, da beide Bewertungsformate den gleichen Erwartungswert für den Anteil an der Maximalpunktzahl besitzen.

Ergebnisse der Punktebewertung

Während im Jahr 2014 mit 93.0% deutlich mehr Prüflinge das gesamte Modul bestanden als noch im Vorjahr, sank wiederum die Bestehensrate der regulären Abschlussklausur sehr deutlich um fast 50% ab, so dass mit 44.8% nicht einmal die Hälfte der Prüflinge diese reguläre Klausur bestanden haben. Dies wirkte sich sehr stark auf die Ergebnisse in den freiwillig zu beantwortenden Zusatzaufgaben aus: Hier hätten nur 24.2% der Prüflinge bei der Bewertung mittels *R4-Scoring* bzw. 16.5% mittels Bonuspunkt-*Scoring* die Zusatzaufgaben bestanden. Dennoch ist damit ein Anstieg der Bestehensrate in den Zusatzaufgaben im Vergleich zum Vorjahr zu verzeichnen, in welchem diese bei nur 10.4% in der *R4*-Gruppe lag.

Um den Zusammenhang der Leistungen in den Zusatzaufgaben und der Leistungen in den regulären Klausuren zu ermitteln, wurden Korrelationen bestimmt. Die Korrelation zwischen dem Gesamtpunktwert (GP) aus Zwischen- und Abschlussklausur und dem Punktwert in den Zusatzaufgaben lag jedoch sowohl beim *R4-Scoring* als auch beim Bonuspunkt-*Scoring* nahe Null und war entsprechend nicht signifikant: $p = .325$

Tabelle 3.6. Ergebnisse für die reguläre Prüfung und die Zusatzaufgaben im Jahr 2014. Dargestellt sind die Ergebnisse für die reguläre Zwischen- (ZK) und Abschlussklausur (AK) sowie deren Summe (Gesamtpunktwert, GP). Weiterhin sind die Ergebnisse der Zusatzaufgaben der beiden Bewertungsformen *R4* und Bonuspunkte (BP) für die Bewertung mit Punkten und für *R4* für die Parameterschätzung nach der Signalentdeckungstheorie (*SDT*) abgedruckt.

	N	Min	Max	*M*	*SD*	Bestehens-rate	*r* mit GP	*r* mit AK
				reguläre Klausuren				
ZK	201	.400	.967	.842	.136	.930	.848***	.301***
AK	201	.233	.900	.561	.111	.448	.761***	1
GP	201	.367	.917	.702	.100	.846	1	.761***
				Zusatzaufgaben, Punkte				
R4	194	.440	.880	.684	.089	.242	.031	.087
BP	194	.473	.840	.666	.078	.165	.023	.053
				Zusatzaufgaben, *SDT*				
R4								
μ_s	194	-.134	2.692	1.010	.494		.318***	.376***
σ_s^2	194	.070	5.845	1.184	.942			
AUC	194	.459	.932	.747	.096		.366***	.408***

Anmerkungen: ***$p < .001$; **$p < .01$; *$p < .05$

für das *R4-Scoring* und $p = .375$ für das Bonuspunkt-*Scoring*. Es wurde daher auf einen Signifikanztest bezüglich des Unterschieds der beiden Korrelationen verzichtet.

Eine Zusammenfassung der Ergebnisse findet sich in Tabelle 3.6. In Abbildung 3.9 sind Streudiagramme der erreichten Gesamtpunkte in den Klausuren im Verhältnis zu den erreichten Punkten in den Zusatzaufgaben für beide Gruppen dargestellt.

Ergebnisse der Signalentdeckungsparameter

Wie in Abschnitt 3.4.2 beschrieben, wurden diejenigen Prüflinge von der Auswertung ausgeschlossen, die bei mehr als fünf Alternativen keine Antwort abgegeben haben.

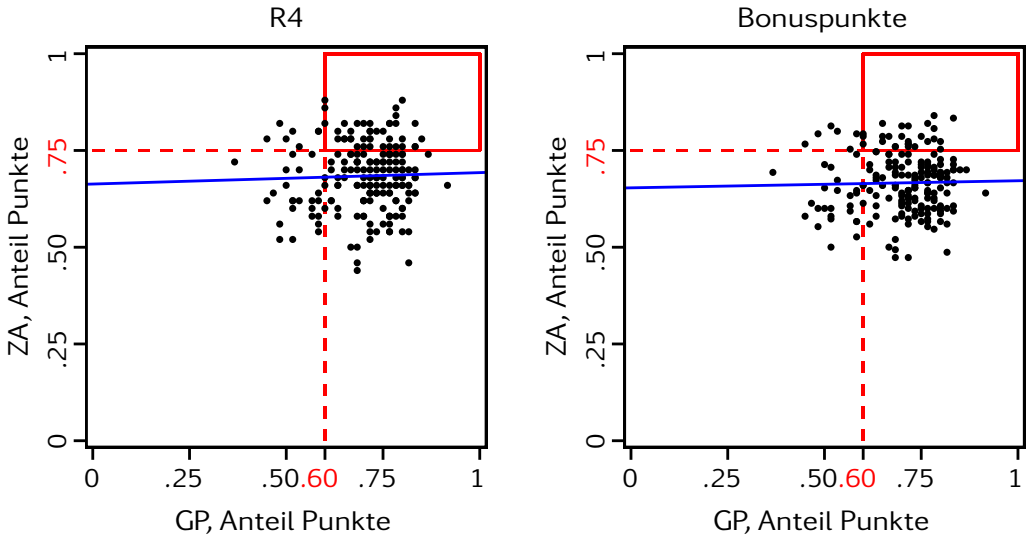

Abbildung 3.9. Gesamtpunkte (GP) in der regulären Prüfung 2014 abgetragen gegen die Punkte in den Zusatzaufgaben (ZA), jeweils als Anteil an der Maximalpunktzahl für die beiden Bewertungsformen *R4* und Bonuspunkte (BP). Jeder Punkt im Koordinatensystem stellt die Leistung eines einzelnen Prüflings dar. Rot markiert sind die für die Aufgabenart maßgeblichen Bestehensgrenzen für 50% Wissen. Prüflinge, die sich auf bzw. innerhalb des roten Rechtsecks befinden, haben sowohl die regulären Klausuren als auch die Zusatzaufgaben bestanden.

Hier ist es jedoch wichtig zu bemerken, dass dies weiterhin dazu führen kann, dass bis zu fünf Antworten für einen Prüfling fehlen. Da das Signalentdeckungsmodell jedoch keine Möglichkeit bietet, fehlenden Datenpunkten Rechnung zu tragen, wurden diese bei der Parameterschätzung verworfen und die Parameter nur an die vorhandenen Daten angepasst, so dass den Schätzungen der einzelnen Prüflinge ggf. leicht unterschiedliche Fallzahlen zugrunde liegen.

Um den Zusammenhang der Signalentdeckungsmaße mit den Leistungen in den regulären Klausuren bestimmen zu können, wurden wiederum Korrelationen für μ_s und *AUC* berechnet. Es zeigte sich, dass sowohl die Korrelation zwischen dem Gesamtpunktwert aus den Klausuren und μ_s mit $r = .318$, $p < .001$, als auch die Korrelation zwischen dem Gesamtpunktwert aus den Klausuren und der *AUC* mit $r = .366$, $p < .001$, auf einem ähnlichen Niveau liegen und signifikant sind. Daraufhin wurden

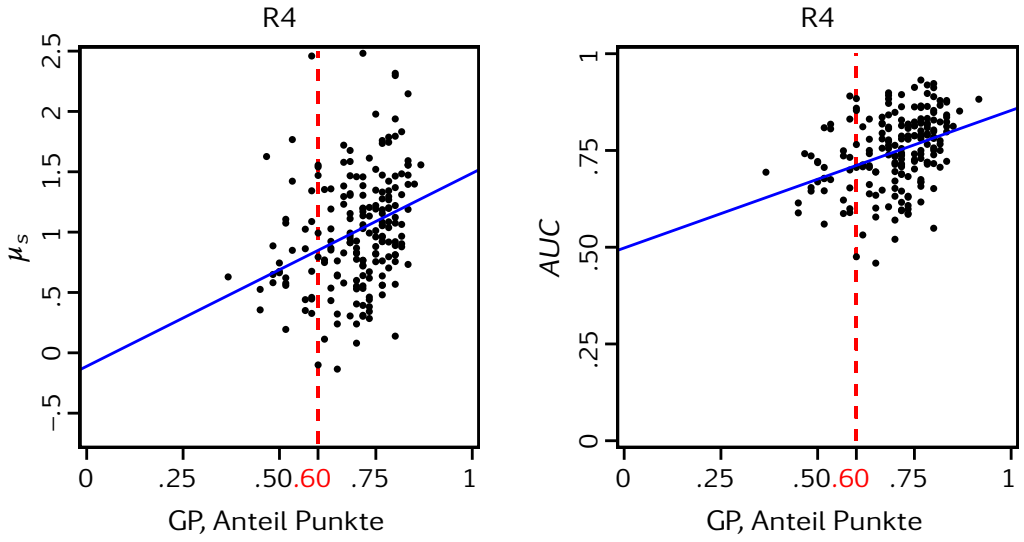

Abbildung 3.10. Anteil der Gesamtpunkte (GP) in der regulären Prüfung 2014 abgetragen gegen die geschätzten Parameter für das Signalentdeckungsmodell in den Zusatzaufgaben. Jeder Punkt im Koordinatensystem stellt die Leistung eines einzelnen Prüflings dar. Rot markiert ist die Bestehensgrenze in der regulären Klausur für 50% Wissen.

die beiden Korrelationen aus der diesjährigen Kohorte mit jenen aus der *R4*-Gruppe der Kohorte von 2013 verglichen. Es konnten jedoch nur numerische, jedoch keine signifikanten Unterschiede nachgewiesen werden: $z = 1.24$, $p = .108$ für μ_s und $z = 1.49$, $p = .068$ für die *AUC*.

Die Ergebnisse für μ_s, σ_s^2 und *AUC* sind auch in Tabelle 3.6 dargestellt. In Abbildung 3.10 sind Streudiagramme der erreichten Gesamtpunkte in den Klausuren im Verhältnis zu den geschätzten Signalentdeckungsparametern in den Zusatzaufgaben dargestellt.

3.4.4 Diskussion

Obwohl die Ergebnisse nicht überwältigend sind, gibt es einige subtile Aspekte, die hier festgehalten werden sollen und zeigen, dass die Information über das Format der Zusatzaufgaben im Vorfeld der Prüfung und die Vergabe von Bonuspunkten zu

höherer Motivation und besserer Leistung der Prüflinge führte.

Zunächst ist festzustellen, dass es im Vergleich zu den Vorjahren vergleichsweise wenige Prüflinge gab, die die Zusatzaufgaben nicht zufriedenstellend bearbeiteten. Insbesondere ist hier erwähnenswert, dass alle Prüflinge die Aufgabenblätter mit den Zusatzaufgaben zurückgaben und auf die korrekte Matrikelnummer achteten. Es mussten nur sieben Prüflinge von der Analyse ausgeschlossen werden, weil diese mehr als fünf Alternativen nicht bearbeitet hatten.

Dennoch ist die Bestehensrate in den Zusatzaufgaben weiterhin sehr niedrig, jedoch bei der gleichermaßen niedrigen Bestehensrate in der regulären Abschlussklausur nicht verwunderlich. Überraschenderweise lässt sich dennoch ein deutlicher und hoch signifikanter Zusammenhang zwischen den geschätzten Parametern des Signalentdeckungsmodells und dem Gesamtpunktwert in den Klausuren feststellen, wohingegen dieser Zusammenhang bei einer klassischen Bewertung mit Punkten völlig verlorengeht.

3.5 Evaluation in 2013 und 2014

3.5.1 Zielstellung

Zunächst sollte im Jahr 2013 eine Einschätzung seitens der Prüflinge gewonnen werden, wie gut sie mit dem *MR*-Format zurechtkamen und inwieweit sie dieses und die drei *MR*-Antwortschlüssel als geeignet für den Einsatz in Prüfungen bzw. in der Lehre halten. Da die Prüflinge im Jahr 2014 bereits einige Monate vor der Klausur auf das *R4*-Format der Zusatzaufgaben durch einen Aushang hingewiesen wurden, sollte überprüft werden, ob dies eine positive Wirkung auf die Einschätzungen der Prüflinge für dieses Format hat.

3.5.2 Methode

Evaluationsbogen

Während der Untersuchungen im Jahr 2013 und 2014 wurde gemeinsam mit den Zusatzaufgaben ein Evaluationsbogen mit zehn Aussagen ausgegeben. Es wurde erfragt, wie gut die Prüflinge mit den gerade eben bearbeiteten Zusatzaufgaben zurecht gekommen sind bzw. wie leicht es ihnen gefallen ist, diese zu beantworten. Weiterhin wurde erfragt, ob die Prüflinge den selbst erlebten Antwortschlüssel generell als sinnvoll für den Einsatz in Prüfungen oder in der Lehre, z.B. in Vorbereitungstests, erachten, ob sie sich mehr solcher Art Aufgaben wünschen und ob ihnen andere, nicht selbst bearbeitete Antwortschlüssel ggf. leichter oder schwerer erscheinen. Die genauen Aussagen sind in Anhang D abgedruckt.

Durchführung

Die Prüflinge sollten im Anschluss an die Bearbeitung der Zusatzaufgaben den Evaluationsbogen ausfüllen. Dafür standen 15 Minuten Bearbeitungszeit zur Verfügung, nach deren Ablauf die Bögen mit den Zusatzaufgaben und die Evaluationsbögen eingesammelt bzw. von den Prüflingen selbst abgegeben wurden.

3.5.3 Ergebnisse der Evaluation

Mittels des Evaluationsbogens sollten die Prüflinge eine Einschätzung ihrer Erfahrungen mit den Zusatzaufgaben rückmelden. Dazu standen fünfstufige *Rating*-Skalen zur Verfügung, wobei große Werte größerer Zustimmung entsprechen. Zur Auswertung wurden je nach Anzahl der Gruppen entweder Varianzanalysen oder t-Tests berechnet, wobei die zugrundeliegenden *Rating*-Skalen als *Perfiat*-Messungen betrach-

Tabelle 3.7. Ergebnisse der Aussagen auf dem Evaluationsbogen über die einzelnen Gruppen für die Evaluation im Anschluss an die Prüfungen in den Jahren 2013 und 2014.

	2013						2014	
	MC		*MTF*		*R4*		*R4*	
Aussage	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
eigene Erfahrungen								
Bearbeitung fiel leicht	2.74	.97	2.13	.82	2.06	.82	2.38	.97
bin gut zurechtgekommen	3.19	1.19	3.00	1.28	2.38	1.17	2.92	1.34
selbst bearbeitetes *MR*-Format								
geeignet für Prüfungen	2.80	1.47	2.23	1.37	1.67	.97	2.14	1.30
geeignet für die Lehre	3.62	1.37	3.38	1.42	2.68	1.43	3.10	1.56
subjektive Schwierigkeit der anderen *MR*-Formate								
MC			3.54	1.22	3.13	1.25	3.57	1.17
MTF	2.94	.98			2.60	1.14	2.90	1.18
R4	3.65	1.28	4.01	1.26				

tet und daher als normalverteilt angenommen wurden (vgl. Bortz & Schuster, 2010).

Ausschnittsweise sind hier die Ergebnisse von sechs der zehn Aussagen berichtet. Deskriptive Daten finden sich in Tabelle 3.7 und Mittelswertsverläufe in Abbildung 3.11. Die verbleibenden vier Aussagen besitzen wenig Aussagekraft über die Zusatzaufgaben und sollen daher außen vor bleiben.

Evaluation in 2013

Zwei Aussagen des Evaluationsbogens bezogen sich direkt auf die gerade beantworteten Zusatzaufgaben. Einerseits sollten die Prüflinge angeben, ob ihnen die Bearbeitung leicht fiel. Hier stimmte die *MC*-Gruppe am meisten zu, gefolgt von der *MTF*- und der *R4*-Gruppe. Dieser Unterschied ist signifikant, $F(2,155) = 10.07$, $p < .001$. Ande-

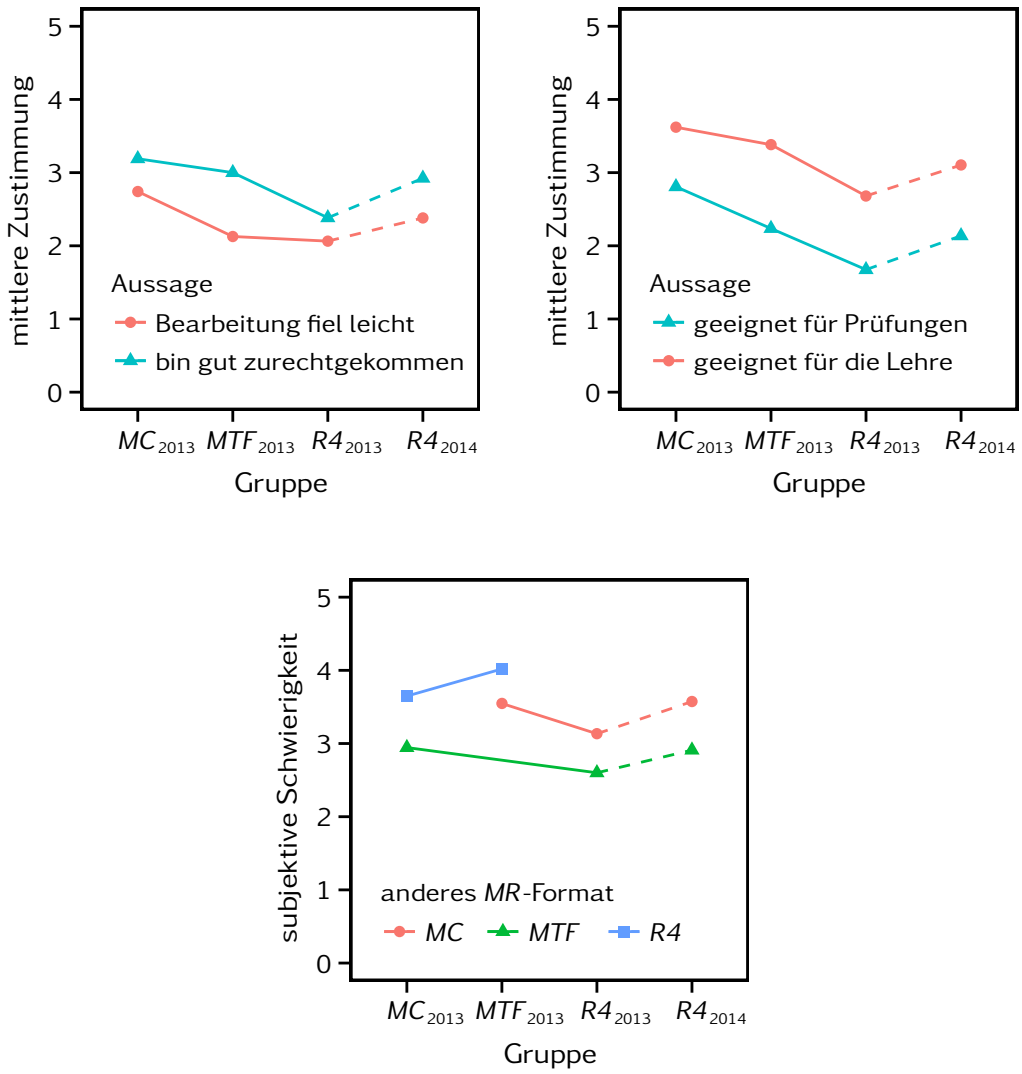

Abbildung 3.11. Mittelwertsverläufe der Aussagen auf dem Evaluationsbogen über die einzelnen Gruppen für die Evaluation im Anschluss an die Prüfungen in den Jahren 2013 und 2014.

rerseits sollten die Prüflinge angeben, ob sie mit dem von ihnen bearbeiteten *MR*-Format zurechtgekommen sind. Es wiederholte sich das Muster der ersten Aussage mit der größten Zustimmung in der *MC*-Gruppe, gefolgt von der *MTF*- und der *R4*-Gruppe. Auch dieser Effekt war signifikant, $F(2, 155) = 6.04$, $p = .003$.

In zwei weiteren Aussagen sollten die Prüflinge einschätzen, ob sie das von ihnen

selbst bearbeitete *MR*-Format der Zusatzaufgaben generell für den Einsatz in Prüfungen bzw. in der Lehre für geeignet halten. Es ergab sich das bereits bekannte Bild der Abfolge der drei Gruppen wie oben. Auch hier waren die Unterschiede zwischen den Gruppen sowohl für den Einsatz in Prüfungen, $F(2,155) = 9.61$, $p < .001$, als auch dem Einsatz in der Lehre, $F(2,157) = 6.09$, $p = .003$, signifikant.

Abschließend sollten sich die Prüflinge vorstellen, wie leicht oder schwer ihnen die Beantwortung der Zusatzaufgaben gefallen wäre, wenn sie nicht das gerade von ihnen selbst bearbeitete *MR*-Format vorgefunden hätten, sondern eines der anderen beiden. Hier zeigen sich in einer unvollständigen zweifaktoriellen Varianzanalyse mit Messwiederholung auf dem *within*-Faktor „anderes *MR*-Format“ und dem *between*-Faktor „eigenes *MR*-Format“ deutliche Effekte. So wird sowohl der Haupteffekt bezüglich des eigenen *MR*-Formats signifikant, $F(2,149) = 10.67$, $p < .001$, als auch der Messwiederholungs-Haupteffekt bezüglich der anderen *MR*-Formate, $F(2,149) = 13.66$, $p < .001$. Die Interaktion verfehlt die Signifikanz, $F(1,149) = 0.84$, $p = .362$.

Evaluation in 2014

Da die Prüflinge im Jahr 2014 bereits einige Zeit vor der Prüfung mittels eines Aushangs auf das in den Zusatzaufgaben verwendete *R4*-Format vorbereitet wurden, sollten die Einschätzungen der Aussagen in diesem Jahr mit denen der *R4*-Gruppe aus dem Jahr 2013 verglichen werden. Es wurde erwartet, dass die 2014er Prüflinge insgesamt besser mit dem *R4*-Format zurechtkommen und ihnen die Bearbeitung leichter fallen sollte. Um dies zu überprüfen, wurden für alle Aussagen Mann-Whitney-*U*-Tests zwischen der *R4*-Gruppe aus dem Jahr 2013 und den Prüflingen aus dem Jahr 2014 berechnet.

Es zeigte sich, dass die Gruppe aus dem Jahr 2014 sowohl angibt, dass ihnen die

Bearbeitung signifikant leichter fiel, $t(229) = 2.06$, $p = .040$, als auch, dass sie mit dem *R4*-Format der Zusatzaufgaben signifikant besser zurechtkamen, $t(227) = 2.53$, $p = .012$. Weiterhin wurde die Eignung des *R4*-Formats für Prüfungen als signifikant besser eingeschätzt, $t(227) = 2.25$, $p = .025$, während jene für den Einsatz in der Lehre zwar deskriptiv auch größer war, dieser Unterschied jedoch nicht signifikant wurde, $t(228) = 1.69$, $p = .093$.

Auch die beiden anderen *MR*-Formate, deren Bearbeitung sich die Prüflinge vorstellen sollten, wurden nun von der Gruppe aus dem Jahr 2014 gegenüber der *R4*-Gruppe aus dem Jahr 2013 als schwerer eingeschätzt, wobei jedoch nur der Unterschied hinsichtlich des *MC*-Formats auch signifikant war, $t(228) = 2.22$, $p = .027$, während der hinsichtlich des *MTF*-Formats nicht signifikant wurde, $t(228) = 1.58$, $p = .115$.

3.5.4 Diskussion

Es sollten zunächst im Jahr 2013 allgemeine Einschätzungen der Prüflinge bezüglich der Bearbeitung und Eignung der Zusatzaufgaben im *MR*-Format eingeholt werden. Hier zeigte sich, dass die Prüflinge recht große Schwierigkeiten mit allen neuen *MR*-Antwortschlüsseln haben, diese jedoch umso größer waren, je weiter sich das *MR*-Format optisch vom gewohnten *SR*-Format entfernt. Insbesondere das *Rating*-Verfahren wurde als schwierig zu durchschauen eingeschätzt.

Dies spiegelt sich auch in der Präferenz der drei Antwortschlüssel wider, wenn die Prüflinge gefragt wurden, als wie schwierig sie die anderen, nicht selbst bearbeiteten Antwortschlüssel einschätzen. Hier war es interessant zu sehen, dass sowohl die *MC*-Gruppe als auch die *MTF*-Gruppe das *R4*-Format als das am schwersten zu bearbeitende Format einschätzen.

Dass das *R4*-Format von den Prüflingen aus dem Jahr 2013 als sehr schwer ange-

sehen wurde, könnte an der völligen Unvertrautheit der Prüflinge mit einem solchen Format gelegen haben. Insbesondere war ihnen nicht klar, wie Antworten in den „eher richtig/falsch“-Kategorien bewertet werden. Daher wurden die Prüflinge im Jahr 2014 bereits ca. drei Monate vor der Prüfung über das Format der Zusatzaufgaben und deren Bewertung informiert (s. Anhang C). Es wurde erwartet, dass hierdurch die Prüflinge besser auf das *R4*-Format, dessen Anforderungen und Bewertung vorbereitet sind. Dies sollte sich in besseren Einschätzungen des *R4*-Formats bezüglich der Bearbeitung durch die Prüflinge äußern. Dies konnte so in den Daten gefunden werden. Es lässt sich daher die klare Empfehlung ableiten, dass die Prüflinge vor der Verwendung eines neuen, für sie unbekannten Aufgabenformats, welcher Art auch immer, zunächst auf dessen spezifische Anforderungen hingewiesen werden sollten.

4

Allgemeine Diskussion

Ziel dieser Arbeit war es einerseits, die Verwendbarkeit von *MR*-Aufgaben mit unterschiedlichen Antwortschlüsseln in Prüfungen zu untersuchen und andererseits die Eignung der Signalentdeckungstheorie für die Bewertung ebenjener Prüfungsaufgaben zu überprüfen. Dazu wurden in mehreren Prüfungen in der Medizin die Leistungen von Prüflingen in den regulären, bestehensrelevanten *SR*-Klausuren mit ihren Leistungen in zusätzlich vorgegebenen *MR*-Aufgaben mit unterschiedlichen Antwortschlüsseln verglichen. Zusätzlich zur traditionellen Bewertung mit Punkten wurden in den Zusatzaufgaben Signalentdeckungsparameter für die einzelnen Prüflinge geschätzt. Weiterhin wurde evaluiert, welche Erfahrungen die Prüflinge mit den Zusatzaufgaben sammelten und wie sie deren Eignung für weitere Prüfungen oder in der Lehre einschätzen.

4.1 Einordnung der Ergebnisse

Um die im vorherigen Kapitel vorgestellten Ergebnisse in ihrer Bedeutung für die vorliegende Arbeit und den gesamten Kontext „Auswertung von Prüfungen mit *multiple response*-Aufgaben“ einordnen zu können, ist es notwendig, sich folgenden Sachverhalt bewusst zu machen: Die vorgestellten Untersuchungen fanden bewusst direkt im universitären Prüfungsalltag statt, so dass direkte Erkenntnisse über das Wissen von „echten“ Prüflingen, gewonnen werden können. Es wurden daher keine zufälligen Versuchspersonen akquiriert, die künstlichen Prüfungsstoff für eine künstliche Prüfung lernen, sondern es wurden im Anschluss an eine reguläre Modulprüfung im Studienfach Medizin den teilnehmenden Studierenden Aufgaben aus demselben Fachgebiet gestellt, auf welches sie sich bereits für die reguläre Klausur vorbereitet hatten.

Leider stellte sich heraus, dass dieser Untersuchungsansatz eigene Probleme mit sich bringt, die großen Einfluss auf die Untersuchungsergebnisse haben, jedoch vorher nicht bedacht wurden. Als wichtigste Erkenntnis, welche alle anderen berichteten Ergebnisse maßgeblich beeinflusst hat, ist festzuhalten, dass viele Prüflinge zum Untersuchungszeitpunkt, entgegen der Erwartung, nicht ausreichend gut auf die Prüfung vorbereitet waren, so dass die Bestehensraten teilweise sehr gering ausfielen.

Es konnte dabei festgestellt werden, dass dieses Ergebnis weder auf die mangelnde Intelligenz der Prüflinge noch auf die besonderes schwierige Gestaltung der Aufgaben zum Untersuchungszeitpunkt zurückzuführen ist. Die erste Aussage lässt sich dadurch belegen, dass die Prüflinge in der Zwischenklausur durchaus dazu in der Lage waren, in ähnlich schwierigen Aufgaben annehmbare Leistungen zu zeigen, so dass die Bestehensraten in den Zwischenklausuren stets hoch waren. Die zweite Aussage kann durch die Auswahl der Aufgaben belegt werden, welche gerade so vorgenom-

men wurde, dass nur in früheren Kohorten bereits verwendete Aufgaben mit entsprechenden teststatistischen Kennwerten Verwendung fanden.

Dennoch fielen die Bestehensraten in den Abschlussklausuren, in deren Rahmen die Untersuchungen durchgeführt wurden, zum Teil dramatisch ab. Hierfür lässt sich jedoch möglicherweise eine relativ einfache Erklärung finden, die in der zeitlichen Struktur des Moduls bzw. des Studiengangs verortet liegt: Die Zwischenklausur fand stets zu einem Zeitpunkt im Semester statt, zu der die Studierenden keine anderen regulären Prüfungen zu absolvieren hatten, während zum Zeitpunkt der Abschlussklausur sieben weitere Prüfungen stattfanden. Den Studierenden standen daher wesentlich mehr zeitliche Ressourcen zur Vorbereitung auf die Zwischenklausur zur Verfügung als zur Vorbereitung auf die Abschlussklausur. Da beide Klausuren in gleicher Weise zum Bestehen des Moduls beitrugen, könnte es eine gängige Strategie der Prüflinge gewesen sein, sich möglichst umfangreich auf die Zwischenklausur vorzubereiten, um bereits in dieser viele Punkte zu sammeln. Demgegenüber wäre für die Abschlussklausur eine geringere Vorbereitung ausreichend, da nur noch wenige fehlende Punkte zum Bestehen benötigt werden.

Die Anwendung dieser ohne Zweifel ökonomischen Strategie kann den einzelnen Prüflingen nicht übel genommen werden. Zwar könnte hier bemängelt werden, dass die Prüflinge nur auf ihren eigenen Vorteil bedacht waren und damit den wissenschaftlichen Grundgedanken „verraten“ hätten bzw., sofern der direkte Nutzen für die eigene Person nicht erkennbar ist, wenig bis keine Leistungsbereitschaft zeigen, was sich bis zur Einführung von Bonuspunkten zur Anrechnung auf das Ergebnis der regulären Klausuren auch stets in hohen *Drop-Out*-Raten geäußert hat. Letztlich muss dabei jedoch bedacht werden, dass das Medizinstudium sehr anspruchsvoll ist und die Hauptaufgabe eines Studierenden, zumindest aus dessen Sicht, darin besteht, dieses

Studium in einer vernünftigen Weise zu beenden. Somit ist klar, dass von den untersuchten Prüflingen nicht erwartetet werden kann, dass sich diese in besonders intensiver Weise auf die Prüfung im Fach Pharmakologie und Toxikologie vorbereiten, wenn sie einerseits bereits einen großen Teil der zum Bestehen notwendigen Punkte zu einem früheren Zeitpunkt erworben haben und andererseits mit einer Vielzahl weiterer Prüfungen konfrontiert sind. Insbesondere auch deshalb, weil die einzelnen Prüflinge persönlich tatsächlich keine Vorteile aus der Untersuchung haben ziehen können.

Somit bleibt letztlich, einzugestehen, dass der Untersuchungsansatz ungünstig gewählt war und in zukünftigen Untersuchungen im direkten Prüfungsalltag auf die eventuellen Besonderheiten des Curriculums besser Rücksicht genommen werden muss. Die ermittelten Ergebnisse sind damit zwar nicht wertlos, leider lassen sie aber keine allgemeingültigen Schlüsse zu. Dies mag jedoch auch von Vorteil sein, bleibt so doch die Möglichkeit bestehen, dass bei Beseitigung der angesprochenen Mängel die Ergebnisse in die erwartete Richtung ausschlagen. Schließlich ist aus anderen Arbeiten aus der Arbeitsgruppe bekannt, dass Prüflinge bei der Beantwortung von bestehensrelevanten Aufgaben, auch wenn diese im *Rating*-Format gestellt werden, durchaus zu ähnlichen Leistungen wie in anderen Aufgabenformaten fähig sind (Much, 2014).

4.2 Das *MR*-Format und die Antwortschlüssel

Die Verwendung von Prüfungsaufgaben im *MR*-Format hat für den Prüfer im Vergleich zum *SR*-Format einige interessante Vorteile. Zum einen ist es möglich, *MR*-Aufgaben flexibler zu gestalten, da es nicht erforderlich ist, eine bestimmte Anzahl von zwar falschen, aber dennoch sinnvollen Alternativen zu generieren. Sollte dies bei einer bestimmten Aufgabenstellung einmal nicht möglich sein, kann diese in einem *SR*-For-

mat nicht verwendet werden, wohingegen bei *MR*-Aufgaben ohne Weiteres mehrere bis hin zu allen Alternativen richtig sein können und somit die Erstellung einer solchen Aufgabe unproblematisch möglich ist. Die oft kritisierte große Ratewahrscheinlichkeit im *MR*-Format lässt sich dabei durch das Modell von Lukas (2015a, 2015b) effektiv behandeln.

Ein weiterer Vorteil des *MR*-Formats stellt die Anforderung an den Prüfling dar, jede einzelne Alternative tatsächlich auf ihre Richtigkeit überprüfen zu müssen, da eine explizite Antwort für jede einzelne Alternative notwendig ist. Dies ist im *SR*-Format nicht gewährleistet, oft gehen Prüfer jedoch davon aus, dass auch falsche Alternativen als solche erkannt worden wären, obwohl darüber tatsächlich keinerlei Daten gewonnen werden.

In der Evaluation der Erfahrungen der Prüflinge mit dem *MR*-Format zeigte sich, dass das *MC*-Format von den drei genutzten Antwortschlüsseln von den Prüflingen am ehesten präferiert wurde. Dies lässt sich eventuell auf seine optische Ähnlichkeit mit dem bekannten *SR*-Format zurückführen. Gleichzeitig zeigte sich jedoch auch, dass die Bestehensrate der *MC*-Gruppe der 2013er Kohorte nur halb so groß war wie jene der beiden anderen Gruppen. Die Präferenz scheint sich daher nicht positiv auf die Leistungen auszuwirken. Möglicherweise könnte die bekannt erscheinende Optik des *MC*-Formats die Prüflinge sogar dazu verleitet haben, eben nicht jede Alternative einzeln auf ihre Richtigkeit zu überprüfen.

An dieser Stelle soll daher insbesondere auf das *MTF*-Format hingewiesen werden. Dieses ist mit den beiden Kategorien „richtig“ und „falsch“ zur Bewertung einer Alternative leicht zu verstehen und beinhaltet gleichzeitig in seiner optischen Erscheinung mit zwei Boxen zu jeder Alternative die Aufforderung, für jede Alternative eine Bewertung anzugeben. Darüber hinaus ist es für den Prüfer in diesem Format im Gegensatz

zum *MC*-Format überhaupt erst möglich, Alternativen zu erkennen, die von Prüflingen ausgelassen oder übersprungen wurden und stellt daher an dieser Stelle ebenfalls eine Verbesserung dar. Leider ist das *MTF*-Aufgabenformat jedoch bisher in keiner der verbreiteten e-Assessment-Plattformen verfügbar.

Das *Rating*-Format wurde von den Prüflingen im Vergleich zu den anderen Formaten als das schwierigste eingeschätzt, was sich möglicherweise mit den ungewohnten Antwortkategorien erklären lässt. Darüber hinaus bedeutet ein Sicherheits*rating* im Bezug auf die eigene Antwort eine größere kognitive Anstrengung als das einfache Beantworten von Aufgaben. Dennoch zeigte sich in der 2013er Kohorte eine ähnliche Bestehensrate wie im *MTF*-Format. Auch die Prüflinge aus der Untersuchung von Much (2014) waren in der Lage, Aufgaben, die im *Rating*-Format gestellt waren, in ähnlich guter Weise zu beantworten, wie jene, die in anderen Formaten präsentiert wurden.

In einer interessanten Arbeit wählten Kampmeyer et al. (2014) einen ähnlichen Ansatz zur Untersuchung der subjektiven Sicherheit von Prüflingen bezüglich ihrer Antwort: Die Autoren präsentierten Medizinstudierenden im dritten und fünften Studienjahr *SR*-Aufgaben aus dem Gebiet der Pharmakologie, die diese beantworten sollten. Im Anschluss an jede Antwort sollten die Prüflinge auf einer vierstufigen Skala, die mit „ich bin mir sicher“ bis „ich rate“ betitelt war, einschätzen, wie sicher sie sich dabei sind. Sie stellten fest, dass die Prüflinge im fünften Studienjahr zwar in ihren Antworten weniger sicher waren, aber dennoch insgesamt öfter korrekt antworteten als die Prüflinge im dritten Studienjahr. Sie zogen daraus den Schluss, dass Studierende im Laufe ihres Studiums mehr Wissen anhäufen, gleichzeitig diesem Wissen jedoch skeptischer gegenüberstehen.

Die Hinterfragung des eigenen Wissens ist ein wichtiger Bestandteil des Studiums,

der durch diese Art der Aufgabenstellung möglicherweise gefördert werden könnte. In einer Prüfung soll jedoch zweifelsfrei der aktuelle Wissensstand festgestellt und bewertet werden, so dass sich ein *Rating*-Format hier vermutlich nicht anbietet. Darüber hinaus wäre die gerichtssichere Bewertung eines solchen Ansatzes zu klären. Für die Vorbereitung einer Prüfung stellt das *Rating*-Verfahren aus den angeführten Gründen jedoch ein wertvolles Werkzeug dar.

4.3 Signalentdeckungstheorie zur Auswertung von Prüfungen

Die Anwendung der Signalentdeckungstheorie für die Auswertung von Prüfungen ist nicht unumstritten. Einerseits ist es hierfür notwendig, neue starke Annahmen zu machen, die sich nur schwer überprüfen lassen und andererseits werden bereits bestehende Annahmen der Signalentdeckungstheorie verletzt. Daneben zeigten sich einige praktische Probleme. Beides soll an dieser Stelle aufgezeigt werden. Es stellt sich die berechtigte Frage, ob die Signalentdeckungstheorie überhaupt ein geeignetes Werkzeug zur Prüfungsauswertung darstellt.

4.3.1 Kritik an der Verwendung der Signalentdeckungstheorie

Verletzung der Voraussetzungen

Eine grundlegende Annahme der Signalentdeckungstheorie, welche hier ohne Zweifel verletzt wird, stellt das dem gesamten Modell zugrundeliegende Zufallsexperiment dar. Hier wird davon ausgegangen, dass genau ein spezifiziertes und unveränderliches Zufallsexperiment mit zwei normalverteilten Zufallsvariablen immer wieder durchgeführt wird. Aus diesen fortwährenden Wiederholungen ergeben sich im Sinne

des Gesetzes der großen Zahl (Poisson, 1837; Seneta, 2013; Tsirelson, 2012) stabile bedingte relative Häufigkeiten der Ereignisse *hit* und *false alarm*, die unter Anwendung des frequentistischen Wahrscheinlichkeitsbegriffs (Fisher, 1925; Friedman, 1999; Kendall, 1949; von Mises, 1928) als Schätzer der zugrundeliegenden Wahrscheinlichkeiten dienen. Dabei wird das Zufallsexperiment durch die Art des immergleichen Signals und des immergleichen Rauschens bestimmt.

Diese Annahme ist für die Auswertung von Prüfungen offensichtlich nicht zu halten, wenn hier die stets unterschiedlichen Alternativen als Signale dienen sollen. Weiterhin ist es fraglich, ob das Rauschen als konstant anzusehen ist oder ob sich dieses gleichfalls nach der Beantwortung einer Alternative verändert, da die Kategorisierung von Alternativen in „richtig“ und „falsch“ den kognitiven Status verändern könnte. Hiermit wäre auch gleichzeitig die Voraussetzung der voneinander unabhängigen Zufallsexperimente verletzt. Darüber hinaus lässt sich die Normalverteilungsannahme bezüglich der beiden Zufallsvariablen nicht prüfen.

Dem ist, auch wenn diese Argumentation das Problem letztlich nicht zu lösen vermag, entgegenzuhalten, dass auch in vielen anderen Anwendungsszenarien der Signalentdeckungstheorie das zugrundeliegende Zufallsexperiment nicht unverändert bleibt. So wird bereits beim Übertrag der Theorie von ihren Ursprüngen in sensorischen Experimenten auf das *recognition*-Paradigma diese Voraussetzung verletzt. Auch in diesem Fall ändert sich der präsentierte Reiz von Durchgang zu Durchgang, ähnlich wie dies in einer Prüfung mit ihren Alternativen der Fall ist. Dennoch erfreut sich die Signalentdeckungstheorie hier großer Beliebtheit und wird zur Auswertung oft gar als Mittel der Wahl empfohlen (Bittrich & Blankenberger, 2011; Macmillan & Creelman, 2010; Swets, 1964; Wixted, 2007). Des Weiteren ist diese Diskrepanz in so gut wie allen Anwendungsfeldern zu finden, sei es bei der Beurteilung von Röntgen-

bildern, bei der Gepäckkontrolle am Flughafen oder der Beurteilung zum Rückfallrisiko von Straftätern. Es scheint daher, dass die Signalentdeckungstheorie hinreichend robust gegen die Verletzung dieser Voraussetzungen ist. Andernfalls wäre an einen Einsatz außerhalb streng experimenteller Szenarien, in denen die Voraussetzungen überprüfbar einzuhalten sind, nicht zu denken. Dieser fundamentalistische Standpunkt würde jedoch bedeuten, dass die Beschäftigung mit der Signalentdeckungstheorie einen rein akademischen Selbstzweck darstellt.

Alternativ ließe sich die *AUC* mit weit weniger starken Annahmen parameterfrei bestimmen (McNicol, 2005). Dazu werden lediglich die ermittelten Paare aus *hit*- und *false alarm*-Raten auf bekannte Weise in einem Koordinatensystem abgetragen und mittels Geraden verbunden. Die *AUC* ergibt sich dann als Fläche unter dem so entstandenen Polygon und lässt sich als Summe der Flächen mehrerer Parallelogramme auf einfache Weise zu berechnen.

Annahme der Eindimensionalität von Prüfungsleistungen

Weiterhin besteht eine Annahme der Signalentdeckungstheorie darin, dass alle Realisierungen des Zufallsexperiments auf einer einzigen Dimension verteilt sind. Hier ist es fraglich, ob dies tatsächlich für alle Aufgaben einer Prüfung zutreffend ist oder ob diesen möglicherweise mehrere unterschiedliche Wissensdimensionen zugrunde liegen. Dies stellt letztlich ein konzeptuelles Problem der Wissensstruktur in einem bestimmten abgeschlossenen Fachgebiet dar. Es stellt sich hierbei völlig unabhängig von der Signalentdeckungstheorie die Frage, ob es verschiedene Wissensdimensionen innerhalb dieses Fachgebiets gibt und falls ja, wie diese strukturiert sind und inwieweit sie sich auf die einzelnen Prüfungsaufgaben auswirken.

So könnte einerseits argumentiert werden, dass das Fach Pharmakologie und To-

xikologie innerhalb der Medizin ein solches abgeschlossenes Wissensgebiet darstellt und sich die gestellten Aufgaben alle auf der Wissensdimension „Pharmakologie und Toxikologie“ lokalisieren lassen. Gleichfalls könnte angenommen werden, dass sich die Wissensdimensionen quer durch alle Gebiete der Medizin ziehen und z.B. die einzelnen Organe abgeschlossene Wissensgebiete bilden. Diese Idee spiegelt sich in der Neuordnung der Prüfungen im Medizinstudiengang der MLU seit dem Sommersemester 2015 wider, welche nun einzelne Klausuren für jedes Organ vorsieht, zu denen die einzelnen Fachgebiete jeweils eine bestimmte Anzahl von Aufgaben beisteuern.

In gleicher Weise ist es vorstellbar, dass erst das gesamte Medizinstudium ein abgeschlossenes Fachgebiet neben anderen Fachgebieten wie Geographie, Anglistik oder Philosophie darstellt. Somit wäre jedwede vorstellbare Frage des Instituts für medizinische und pharmazeutische Prüfungsfragen (IMPP, 2015) auf der gleichen Dimension lokalisiert.

Fehlende Beachtung der unterschiedlichen Schwierigkeiten von Alternativen

Ein weiteres Problem bei der Verwendung der Signalentdeckungstheorie zur Prüfungsauswertung ergibt sich direkt aus der Annahme eines immergleichen Zufallsexperiments. Somit wirkt jedes einzelne präsentierte Signal in gleicher Weise auf die Parameterschätzung ein. Dies ist in Anbetracht von unveränderlichen visuellen oder akustischen Reizen sicherlich sinnvoll, jedoch ergibt sich für die Auswertung von Prüfungen mit einer Vielzahl unterschiedlicher Alternativen das Problem, dass hierdurch jede einzelne Alternative durch das Modell als gleich schwer angesehen wird, ohne dass es eine Möglichkeit gäbe, zwischen mehr oder weniger schweren Alternativen zu differenzieren. Eine Lösung hierfür stellt eine Erweiterung des Signalentdeckungsmodells um *Item*effekte dar, wie sie z.B. DeCarlo (2011) beschreibt.

Ähnliche Probleme mit dem klassischen Bewertungsverfahren

Nachdem nun auf einige kritische Punkte und mögliche Lösungen bei der Verwendung der Signalentdeckungstheorie zur Prüfungsauswertung eingegangen wurde, soll hier nicht unerwähnt bleiben, dass auch die klassische Methode, einen Summens*core* zu berechnen, zum Teil die gleichen Probleme aufweist. Auch hier werden zum Teil starke Annahmen gemacht, die jedoch oft ungeprüft und undiskutiert bleiben.

So ist es oft üblich, alle korrekt beantworteten Alternativen einer Prüfung mit der gleichen Punktzahl zu bewerten, so dass diese, ähnlich wie jedes einzelne Signal, in gleicher Weise zum Summens*core* beiträgt. Bei diesem Vorgehen ist also keine Unterscheidung in leichte bzw. schwere Alternativen möglich. Dem kann entgegengewirkt werden, indem unterschiedlich viele Punkte für verschiedene Alternativen vergeben werden, jedoch ist deren Größe meist willkürlich und steht ggf. in keinem Zusammenhang zur tatsächlichen Schwierigkeit.

Weiterhin stellt ein Summens*core*, genau wie die Schätzung der Signalentdeckungsparameter, die Aggregation von Daten zu einem einzigen Wert dar, dessen Lage sich auf nur einer einzigen Dimension bewegt. Dies ist letztlich dem gesetzlich geregelten Prüfungssystem zuzuschreiben, das die Bewertung einer Prüfungsleistung mittels einer überschaubaren Anzahl Noten erforderlich macht, welche wiederum aus dem Summens*core* abgeleitet werden.

4.3.2 Praktische Probleme bei der Prüfungsauswertung mittels der Signalentdeckungstheorie

Neben den bereits weiter oben behandelten theoretischen Problemen bei der Verwendung der Signalentdeckungstheorie zur Prüfungsauswertung, haben sich bei der Er-

stellung dieser Arbeit im wesentlichen zwei praktische Probleme ergeben: Einerseits, wie mit fehlenden Daten umgegangen werden kann bzw. muss und andererseits, wie aus einer erfolgten Parameterschätzung eine Note abgeleitet werden kann.

Fehlende Daten

Zunächst soll hier das Problem der fehlenden Daten diskutiert werden. Diese ergeben sich immer dann, wenn Prüflinge eine Alternative nicht bearbeiten, wobei es hierfür eine Vielzahl von Gründen geben kann. So ist es möglich, dass ein Prüfling eine Alternative bzw. deren Antwortboxen zwischen der Vielzahl der Antwortboxen der anderen Alternativen übersehen hat. Weiter ist es möglich, dass die für die Prüfung zur Verfügung stehende Zeit nicht ausgereicht hat und so die letzten Alternativen gar nicht bearbeitet werden konnten. Oder es ist möglich, dass ein Prüfling die richtige Antwort für eine Alternative nicht kennt und keine Antwort gibt, statt zu raten. Es ist damit klar, dass fehlende Daten bei *MR*-Aufgaben überhaupt nur bei den Antwortschlüsseln *MTF* und *Rx* identifizierbar sind, da dort tatsächlich eine ganze Zeile von Antwortboxen frei bleibt. Bei der *MC*-Variante ist jedoch bei einer freien Box nicht unterscheidbar, ob dies eine tatsächliche „falsch“-Antwort darstellt oder eine nicht bearbeitete Alternative.

Bei der Bewertung mittels eines klassischen Summenwerts würden fehlende Daten als inkorrekt gewertet und mit Null Punkten bewertet werden. Damit reduziert sich der maximal erreichbare Anteil von Punkten an der Maximalpunktzahl und eine schlechtere Note ist die Folge. Während dieses Vorgehen bei Prüflingen, die die gestellten Aufgaben in der zur Verfügung stehenden Zeit nicht bearbeiten konnten, gerechtfertigt ist, werden Prüflinge, die eine Alternative ggf. einfach übersehen haben, benachteiligt.

Bei einer Auswertung mit der Signalentdeckungstheorie stellen fehlende Daten ein weitaus größeres Problem dar, da sie an keiner Stelle während der Parameterschät-

zung berücksichtigt werden können. Daher wurden bei allen in dieser Arbeit berichteten Parameterschätzungen jeweils alle fehlenden Daten vor der Auswertung gestrichen und es gingen nur die Bewertungen von Alternativen in die Auswertung ein[8]. Dieses Vorgehen hat den Vorteil, dass nun unsystematisch fehlende Daten, wie übersehene Alternativen, die geschätzten Parameter nicht negativ beeinflussen und so ein genaueres Leistungsbild entsteht. Problematisch ist jedoch, dass Alternativen, die von Prüflingen bewusst ausgelassen wurden, z.B. weil sie die richtigen Antworten nicht kannten und ggf. später noch einmal zu diesen Alternativen zurückkehren wollten, in gleicher Weise unberücksichtigt bleiben, obwohl hier eindeutig Nicht-Wissen vorliegt und die tatsächliche Leistung überschätzt wird. Eine Trennung zwischen bewusst und aus Versehen ausgelassenen Alternativen ist seitens des Prüfers jedoch unmöglich. Es stellt sich daher die Frage nach einem geeigneten Korrekturmechanismus für fehlende Daten, wenn die Signalentdeckungstheorie als sinnvoll verwendbares Werkzeug zur Auswertung von Prüfungen etabliert werden soll. Hierzu ist weitere Forschung angeraten.

Bestimmung einer Note

Sofern sich eine Lösung für die Schätzung der Signalentdeckungsparameter in Anbetracht der gerade geschilderten Probleme finden lässt, erfordert der Prüfungsprozess an seinem Ende die Vergabe einer Note für die erreichte Leistung. Hier stellt sich nun die Frage, welches der ermittelten Leistungsmaße dazu geeignet ist, in eine Note überführt zu werden und auf welche Weise dies geschehen soll. Es wurde bereits in Kapitel 2 darauf hingewiesen, dass die Fläche unter der *ROC*-Kurve, die *AUC*, hierfür

[8] Hierbei ist zu beachten, dass bereits vor der Auswertung alle Prüflinge von der weiteren Analyse ausgeschlossen wurden, die mehr als fünf Alternativen nicht bearbeitet hatten. Somit gingen in jede Parameterschätzung die Antworten auf mindestens 45 von maximal 50 Alternativen ein.

am geeignetsten erscheint, da sich diese als die Wahrscheinlichkeit für eine richtige Antwort in einem *2AFC*-Paradigma interpretieren lässt (Macmillan & Creelman, 2010; McNicol, 2005; Wickens, 2002), also im Prüfungskontext, wenn eine richtige und eine falsche Alternative gleichzeitig vorgegeben werden und ein Prüfling angeben soll, welche der beiden richtig ist.

Der Wertebereich der *AUC* liegt somit im abgeschlossenen Einheitsintervall ($0 \leq AUC \leq 1$), wobei ein Wert von $AUC = 1$ eine perfekte Leistung darstellt, ein Wert von $AUC = .50$ dem Rateniveau entspricht und ein Wert von $AUC = 0$ systematischen entgegengesetzten Antworten entspricht, also z.B. in einem *MTF*-Format alle richtigen Alternativen mit „falsch“ und alle falschen Alternativen mit „richtig“ beantwortet werden.

Es ist daher naheliegend, zur Notenfindung einen ähnlichen Ansatz wie im Wahrscheinlichkeitsmodell von Lukas (2015a, 2015b, s. Abschnitt 3.1.4) zu wählen: In einem *2AFC*-Paradigma beträgt die Ratewahrscheinlichkeit $p_R = 1/2 = .50$. Somit ergibt sich für das 50%-Kriterium ($p_W = .50$) nach der Modellgleichung 3.1 eine Wahrscheinlichkeit für eine korrekte Antwort von $p_k = .75$. Die Bestehensgrenze sollte daher auf den Wert $AUC = .75$ festgelegt werden, da auch hier bei Überschreitung dieses Werts mehr Alternativen gewusst als nicht gewusst werden.

4.3.3 Möglichkeiten zur Verbesserung des Verfahrens

Aus der Literatur ist bekannt, dass das *Yes*/*No*-Paradigma dem *2AFC*-Verfahren zur Bestimmung der Sensitivität einer Versuchsperson unterlegen ist (Blackwell, 1953; Fechner, 1860; Jones, 1956; Katkov, Tsodyks & Sagi, 2006; Tyler & Chen, 2000). Sollen also Signalentdeckungsparameter bestimmt werden, ist es daher sinnvoll, stattdessen auf das *2AFC*-Verfahren zurückzugreifen. Dies setzt jedoch die Möglichkeit

voraus, Prüflingen jeweils eine richtige und eine falsche Alternative in einer Aufgabe präsentieren zu können, von denen die richtige Alternative markiert werden soll.

Für Klausuren, die auf Papier realisiert werden, ist dies zwar denkbar, setzt jedoch einigen Aufwand bei der Formatierung der Klausurbögen voraus. Darüber hinaus könnte es hier jedoch möglicherweise zu Quereffekten zwischen einzelnen Aufgaben kommen, da es Prüflingen im Normalfall möglich ist, zu bereits bearbeiteten Aufgaben zurückzukehren. Es wäre daher besondere Sorgfalt bei der Auswahl der Aufgaben vonnöten.

Eine Lösung für dieses Problem könnten daher die immer weiter verbreiteten e-Assessment-Plattformen darstellen. Hier wäre es möglich, eine Rückkehr zu bereits bearbeiteten Aufgaben zu unterbinden, so dass keine Quereffekte entstehen. Des Weiteren ließe sich hier die Präsentation der beiden Alternativen leichter verwirklichen und sogar bis zu einem gewissen Grad zufällige Paarungen von richtigen und falschen Alternativen realisieren. Allerdings bietet bisher keine der verbreiteten Plattformen dies als Aufgabenformat an.

4.4 Fazit und Ausblick

Ungeachtet der offensichtlichen Probleme, die sich bei der Anfertigung der vorliegenden Arbeit ergeben haben und bereits weiter oben diskutiert wurden, bleibt festzuhalten, dass sowohl das *MR*-Aufgabenformat prinzipiell für den Einsatz in Prüfungen geeignet ist als auch, dass die Signalentdeckungstheorie unter bestimmten Umständen ein geeignetes Auswertungsverfahren für Prüfungen darstellen kann.

Es ist empfehlenswert, den Prüflingen im Vorfeld der Prüfung Gelegenheit zu geben, sich mit dem Format vertraut zu machen. Eine Möglichkeit dazu können semesterbegleitend stattfindende vorbereitende Tests mit ähnlichen Aufgaben darstellen. Insbe-

sondere sollte während der Vorbereitungszeit auf die besonderen Herausforderungen des *MR*-Formats, wie mehrere richtige Alternativen, eingegangen werden. Auch die Art und Weise der Bewertung sollte mitgeteilt werden, so dass sich die Prüflinge darüber im Klaren sind, was von ihnen erwartet wird.

Letztlich schuldig bleibt diese Arbeit eine Antwort auf die Frage, ob die geschätzten Signalentdeckungsparameter die Leistungen von Prüflingen besser widerspiegeln als das klassische Verfahrens mittels Summe*score*. Dies liegt auf der einen Seite an den unzureichend belastbaren Daten, auf der anderen Seite fehlt jedoch auch ein zuverlässiges Außenkriterium, mit dem die ermittelten Leistungen verglichen werden könnten. Hier gilt es daher, einerseits die Quelle der Daten zu stabilisieren und dazu ggf. auf klassische experimentelle Ansätze (z.B. Diedenhofen & Musch, 2015; Willing, Ostapczuk & Musch, 2015) oder bestehensrelevante Aufgaben zurückzugreifen (Much, 2014). Andererseits ist die Eignung von Außenkriterien wie Intelligenz, Studien- oder Berufserfolg zu diskutieren (Abele, Bargel, Pajarinen & Schmidt, 2009; Hülsheger, Maier & Stumpp, 2007; Kramer, 2009).

A

R-Script

Schätzung der Parameter eines Signalentdeckungsmodells

```
1  ######################################################################
2  ##                                                                  ##
3  ##  Funktion zur Parameterschätzung für das Signalentdeckungs-      ##
4  ##  modell                                                          ##
5  ##                                                                  ##
6  ##  Funktionsargumente:                                             ##
7  ##    - Signal:   numerischer Vektor                                ##
8  ##                  mit 0 für noise-Trial und                       ##
9  ##                      1 für Signal-Trial                          ##
10 ##    - Category: numerischer Vektor mit Antworten der VP           ##
11 ##                  hohe Werte stehen für größere Sicherheit,       ##
12 ##                  dass ein Signal vorliegt)                       ##
13 ##    - Binary:   bool'scher Vektor                                 ##
14 ##                  TRUE:  Category enthält nur Nullen und          ##
15 ##                         Einsen,                                  ##
16 ##                         entspricht equal-variance-Modell,        ##
17 ##                         keine Schätzung der Signal-Varianz       ##
18 ##                  FALSE: Category-Daten entstammen einem          ##
19 ##                         Rating-Experiment,                       ##
20 ##                         Schätzung der Signal-Varianz             ##
21 ##                                                                  ##
22 ##                                                                  ##
```

```
##    - correctExtremes: Boolean                                    ##
##                       TRUE:  Korrektur der relativen             ##
##                              Häufigkeiten nach dem log-          ##
##                              linearen Ansatz                     ##
##                       FALSE: keine Korrektur der relativen       ##
##                              Häufigkeiten                        ##
##    - highestCategory: Integer                                    ##
##                       nötig, wenn correctExtremes = TRUE,        ##
##                       gibt die höchste der VP zur Verfügung      ##
##                       stehende Kategorie an, so dass alle        ##
##                       Kategorien korrigiert werden können        ##
#####################################################################

##  benötigte R-Packages laden                                  ####
##  (müssen bereits installiert sein)                           ####
require(sensR)
require(ordinal)


SDT <- function(
         Signal, Category, Binary = T,
         correctExtremes = T, highestCategory)
{
  ##  leere Rückgabe-Parameter erzeugen                         ####
  params        <- data.frame(matrix(ncol = 8, nrow = 1))
  names(params) <- c("dprime", "SD", "AUC", "ChiSq",
                     "p", "df", "logLik", "Model")


  ##  NA-Werte in Category behandeln                            ####
  Signal <- Signal[!is.na(Category)]
  Category <- Category[!is.na(Category)]


  ##  Extremwertkorrektur vornehmen                             ####
  if(correctExtremes)
  {
    if(Binary)
    {
      usedCategories <- c(0,1)
    } else
    {
      usedCategories <- c(1:highestCategory)
    }
  }

  nCategories <- length(usedCategories)

  frequencies <- NULL
  for (i in 1:nCategories)
```

```
73  {
74    frequencies[i] <-
75      length(Category[Signal == 0 & Category == usedCategories[i]])
76    frequencies[nCategories + i] <-
77      length(Category[Signal == 1 & Category == usedCategories[i]])
78  }
79
80  nTotalObs     <- sum(frequencies)
81  nSignalTrials <- sum(Signal)
82  nNoiseTrials  <- length(Signal) - nSignalTrials
83
84  if(correctExtremes)
85  {
86    frequencies <- frequencies + .5
87
88    nSignalTrials <- nSignalTrials + 2
89    nNoiseTrials  <- nNoiseTrials + 2
90
91    frequencies[1:nCategories] <-
92      frequencies[1:nCategories] /
93        sum(frequencies[1:nCategories]) * nSignalTrials
94
95    frequencies[(nCategories + 1):length(frequencies)] <-
96      frequencies[(nCategories + 1):length(frequencies)] /
97        sum((nCategories + 1):length(frequencies)) * nNoiseTrials
98  }
99
100 Data <- data.frame(
101   c(rep(0, nCategories), rep(1, nCategories)),
102   as.factor(rep(usedCategories, 2)),
103   frequencies
104 )
105 names(Data) <- c("Signal", "Category", "Frequency")
106
107
108 ##  Parameterschätzung durchführen                              ####
109 if(nlevels(Data$Category) >= 2)
110 {
111   try(
112     if(Binary)
113     {
114       Estimates <- ordinal::clm(
115           Data$Category ~ Data$Signal,
116           weights = Data$Frequency,
117           link = "probit"
118       )
119     } else
120     {
121       Estimates <- ordinal::clm(
122           Data$Category ~ Data$Signal,
```

```
123             ~Data$Signal,
124             weights = Data$Frequency,
125             link = "probit"
126           )
127       }, silent = T
128     )
129
130     if(exists("Estimates"))
131     {
132       dprime <- Estimates$beta
133       SD     <- ifelse(Binary, 1, exp(Estimates$zeta))
134       AUC    <- sensR::AUC(dprime, scale=SD)[[1]]
135
136       nNoiseObs  <- sum(Data[Data$Signal == 0,]$Frequency)
137       nSignalObs <- sum(Data[Data$Signal == 1,]$Frequency)
138
139       fittedFrequencies <- fitted(Estimates)
140       fittedFrequencies[1:nCategories] <-
141         fittedFrequencies[1:nCategories] * nNoiseObs
142       fittedFrequencies[(nCategories + 1):(2 * nCategories)] <-
143         fittedFrequencies[(nCategories + 1):(2 * nCategories)] *
144           nSignalObs
145
146       ChiSq <- sum(((frequencies - fittedFrequencies)^2) /
147                      fittedFrequencies)
148       df    <- 2*(nCategories - 1) - Estimates$edf
149       suppressWarnings(p <- pchisq(ChiSq, df, lower.tail = F))
150
151       logLik <- Estimates$logLik
152
153       params <- data.frame(dprime, SD, AUC, ChiSq,
154                            p, df, logLik, row.names = NULL)
155       params
156       return(params)
157     } else
158     {
159       params
160       return(params)
161     }
162   } else
163   {
164     warning("Less than two Categories were used.
165              Model cannot converge.")
166     params
167     return(params)
168   }
169
170 }
171
172 ####  End of File  ##################################################
```

B

Verzeichnis der Zusatzaufgaben

Im Folgenden findet sich eine Auflistung aller Zusatzaufgaben, aufgeteilt nach dem Jahr ihrer Verwendung. Sowohl die Zusatzaufgaben als auch die jeweiligen Alternativen sind in der tatsächlich genutzten Reihenfolge wiedergegeben, die Zusatzaufgaben sind zusätzlich entsprechend nummeriert. Diejenigen Alternativen, denen ein „x“ vorangestellt ist, stellen die tatsächlich richtigen Alternativen dar. Dies dient nur zur Information des Lesers dieser Arbeit und war den Prüflingen selbstverständlich unbekannt.

B.1 Zusatzaufgaben 2012 im *SR*-Format

1	Welches der folgenden Antimykotika sollte am ehesten zur Therapie systemischer Pilzinfektionen verwendet werden?	
	×	Amphotericin B
		Flucytosin
		Ketoconazol
		Fluconazol
		Griseofulvin

2	Eine im 8. Monat schwangere Frau kommt zu Ihnen in Ihre Praxis. Die Untersuchung ergibt einen Harnwegsinfekt mit E. coli. Welches Medikament würden Sie vorzugsweise geben, ohne ein Risiko für den Fötus einzugehen?	
	×	Cefadroxil (orales Cephalosporin der ersten Generation)
		Cotrimoxazol
		Penicillin V
		Ofloxacin
		Tetracyclin

3	Ein Vater kommt mit seinem kleinen Sohn, weil der akut krank sei. Man findet Symptome des Gastrointestinal-Traktes, Kopfschmerzen und Übelkeit. Das Kind ist lethargisch, hat erhöhte Körpertemperatur und die Augenspiegelung zeigt hellrote Retinavenen. Der behandelnde Arzt entnimmt Blut für Laboruntersuchungen. Wenn man annimmt, daß dieses Kind an einer Vergiftung leidet, welches ist die wahrscheinlichste Ursache?	
		Ein Inhibitor der Cholinesterase
	×	Kohlenmonoxid
		Ethylenglykol (Frostschutzmittel)
		Blausäure
		Schwefeldioxid

4		Welche Aussage zu der Vergiftung und dem Toxin bei dem Kind ist richtig? (GEHÖRT ZUR FRAGE DAVOR)
		Bläuliche Hautfärbung tritt bei 80% der Patienten mit dieser Vergiftung auf.
	×	Einatmen von Feuerrauch kann diese Vergiftung hervorrufen.
		Die Behandlung umschließt unter anderen die Gabe von Atropin und Pralidoxim.
		Therapie besteht in der Gabe von Fomepizol.
		Sauerstoff sollte nicht gegeben werden, bevor die Analyse des Carboxyhämoglobingehaltes abgeschlossen ist.

5		Welche Antwort trifft zu? Das Antiarrhythmikum Lidocain
		verlängert die Dauer des monophasischen Aktionspotentiales.
		hat einen positiv inotropen Effekt.
		führt zur Hyperpolarisierung der Myocyten.
		verlängert beim Patienten die Dauer des QT Intervalls im Oberflächen-EKG.
	×	reduziert die Erregbarkeit von Myocyten im Ventrikel.

6		Welcher der folgenden Patienten wird sehr wahrscheinlich davon profitieren, wenn man ihn/sie mit intravenösem Glucagon behandelt?
		Eine 18-jährige Frau, die eine Überdosis Kokain geschluckt hat und nun einen Blutdruck von 190/100 mmHg zeigt.
		Eine 27-jährige Frau mit schwerem Durchfall bedingt durch Morbus Crohn.
		Eine 57-jährige Frau mit Typ II Diabetes, die ihre Tabletten Glibenclamid in den letzten drei Tagen nicht genommen hat.
	×	Ein 62-jähriger Mann mit schwerer Bradykardie und Hypotonus hervorgerufen durch eine Überdosierung von Atenolol.
		Ein 74-jähriger Mann mit Laktazidose als Folge einer schweren Infektion und Schock.

7	Genetischer Polymorphismus im Metabolismus ist eine anerkannte Ursache in der Variabilität der analgetischen Wirkung von	
		Buprenorphin.
	×	Codein.
		Methadon.
		Morphin.
		Proproxyphen.

8	Welche der folgenden Substanzen führt am wahrscheinlichsten bei Überdosierung zu folgenden Symptomen: massiver Blutdruckabfall, Krampfanfälle und Herzrhythmusstörungen?	
		Paracetamol
		Diazepam
		Ethylenglykol
		Morphin
	×	Amitryptilin

9	Was ist eine typische unerwünschte Wirkung von Colchicin?	
		Psychische Störungen
		Blutdruckanstieg
		Rötliche Hautausschläge
	×	Starke Durchfälle
		Plötzliches gastrointestinales Bluten

10	Welche Aussage trifft zu? Männer, die hohe Dosen von anabolen Steroiden zu sich nehmen, haben ein erhöhtes Risiko für	
		Anämie
	×	cholestatischen Ikterus und Anstieg von Leberenzymen
		Hirsutismus
		Hyperprolactinämie
		Vergrößerung der Testikel

B.2 Zusatzaufgaben 2012 im *MR*-Format

1	Welche der folgenden Antimykotika können zur Therapie systemischer Pilzinfektionen verwendet werden?	
	×	Amphotericin B
	×	Nystatin
	×	Tolnaftat
	×	Amorolfin
		Griseofulvin

2	Eine im 8. Monat schwangere Frau kommt zu Ihnen in Ihre Praxis. Die Untersuchung ergibt einen Harnwegsinfekt mit E.coli. Welches Medikament würden Sie vorzugsweise geben, ohne ein Risiko für den Fötus einzugehen?	
	×	Cefadroxil (orales Cephalosporin der ersten Generation)
		Cotrimoxazol
		Penicillin V
		Ofloxacin
		Tetracyclin

3	Ein Vater kommt mit seinem kleinen Sohn, weil der akut krank sei. Man findet Symptome des Gastrointestinal-Traktes, Kopfschmerzen und Übelkeit. Das Kind ist lethargisch, hat erhöhte Körpertemperatur und die Augenspiegelung zeigt hellrote Retinavenen. Der behandelnde Arzt entnimmt Blut für Laboruntersuchungen. Wenn man annimmt, dass dieses Kind an einer Vergiftung leidet, welches ist die wahrscheinlichste Ursache?	
		Ein Inhibitor der Cholinesterase
	×	Kohlenmonoxid
		Ethylenglykol (Frostschutzmittel)
		Blausäure
		Schwefeldioxid

4	Welche Aussagen zu der Vergiftung und dem Toxin bei dem Kind sind richtig? (GEHÖRT ZUR FRAGE DAVOR)
	Bläuliche Hautfärbung tritt bei 80% der Patienten mit dieser Vergiftung auf.
×	Einatmen von Feuerrauch kann diese Vergiftung hervorrufen.
	Die Behandlung umschließt unter anderen die Gabe von Atropin und Pralidoxim.
	Therapie besteht in der Gabe von Fomepizol.
	Sauerstoff sollte nicht gegeben werden, bevor die Analyse des Carboxyhämoglobingehaltes abgeschlossen ist.

5	Das Antiarrhythmikum Lidocain
	verlängert die Dauer des monophasischen Aktionspotentiales.
	hat einen positiv inotropen Effekt.
	führt zur Hyperpolarisierung der Myocyten.
	verlängert beim Patienten die Dauer des QT Intervalls im Oberflächen-EKG.
×	reduziert die Erregbarkeit von Myocyten im Ventrikel.

6	Welche der folgenden Patienten werden sehr wahrscheinlich davon profitieren, wenn man ihn/sie mit intravenösem Glucagon behandelt?
	Eine 18-jährige Frau, die eine Überdosis Kokain geschluckt hat und nun einen Blutdruck von 190/100 mmHg zeigt.
	Eine 27-jährige Frau mit schwerem Durchfall bedingt durch Morbus Crohn.
	Eine 57-jährige Frau mit Typ II Diabetes, die ihre Tabletten Glibenclamid in den letzten drei Tagen nicht genommen hat.
×	Ein 62-jähriger Mann mit schwerer Bradykardie und Hypotonus hervorgerufen durch eine Überdosierung von Atenolol.
	Ein 74-jähriger Mann mit Laktazidose als Folge einer schweren Infektion und Schock.

7	Genetischer Polymorphismus im Metabolismus ist eine anerkannte Ursache in der Variabilität der analgetischen Wirkung von	
		Buprenorphin.
	×	Codein.
		Methadon.
		Morphin.
		Proproxyphen.

8	Welche der folgenden Substanzen führt am wahrscheinlichsten bei Überdosierung zu folgenden Symptomen: massiver Blutdruckabfall, Krampfanfälle und Herzrhythmusstörungen?	
		Paracetamol
		Diazepam
		Ethylenglykol
		Morphin
	×	Amitryptilin

9	Was ist eine typische unerwünschte Wirkung von Colchicin?	
		Psychische Störungen
		Blutdruckanstieg
		Rötliche Hautausschläge
	×	Starke Durchfälle
		Plötzliches gastrointestinales Bluten

10	Männer, die hohe Dosen von anabolen Steroiden zu sich nehmen, haben ein erhöhtes Risiko für	
		Anämie
	×	cholestatischen Ikterus und Anstieg von Leberenzymen
		Hirsutismus
		Hyperprolactinämie
		Vergrößerung der Testikel

B.3 Zusatzaufgaben 2013

Im Jahr 2013 wurden als Zusatzaufgaben im Gegensatz zu 2012 nur Aufgaben im *MR*-Format verwendet. Im Jahr 2013 wurden dabei die Antwortschlüssel *MC, MTF* und *R4* benutzt (vgl. Abschnitt 3.3.2 sowie Abbildungen 1.1 und 3.6).

1	Welche der folgenden Substanzen ist sinnvoll für die Behandlung der Myasthenia gravis?	
		Diazepam
		Atropin
	×	Neostigmin
		Scopolamin
	×	Pyridostigmin

2	Welche Vorteile bietet Reteplase gegenüber Streptokinase?	
	×	Reteplase bedingt seltener Schüttelfrost.
	×	Reteplase bedingt seltener Kopfschmerzen.
	×	Reteplase wird von Fibrin aktiviert.
	×	Reteplase bedingt seltener Gelenkschmerzen.
	×	Reteplase kann wiederholt gegeben werden.

3	Welche der folgenden antiviralen Substanzen zeigen die größte selektive Wirkung auf ein Virus?	
		Interferon
	×	Aciclovir
		Amantadin
		Zidovudin
		Ribavirin

4	Welche Antworten treffen zu? Das Antiarrhythmikum Lidocain
	verlängert die Dauer des monophasischen Aktionspotentiales.
×	hat einen negativ inotropen Effekt.
×	verzögert die Repolarisationsgeschwindigkeit.
	verlängert beim Patienten die Dauer des QT Intervalls im Oberflächen-EKG.
×	reduziert die Erregbarkeit von Myocyten im Ventrikel.

5	Welche der folgenden Patienten werden sehr wahrscheinlich davon profitieren, wenn man sie mit intravenösem Glucagon behandelt?
×	Eine 18-jährige Frau bei der man zur Diagnose eine Erschlaffung des oberen Teils des Magen-Darm-Traktes erreichen will.
	Eine 27-jährige Frau mit schwerem Durchfall bedingt durch Morbus Crohn.
×	Eine 57-jährige Frau mit Typ II Diabetes zur Behandlung eins hypoglykämischen Schocks.
×	Ein 62-jähriger Mann mit schwerer Bradykardie und Hypotonus hervorgerufen durch eine Überdosierung von Atenolol.
	Ein 74-jähriger Mann mit Laktazidose als Folge einer schweren Infektion und Schock.

6	Welche Aussagen zu Oseltamivir treffen am ehesten zu?
×	Es hemmt ein Protein, das für die Ablösung der Viren von der Zelle wichtig ist.
	Es hemmt die virale DNA-Replikation.
×	Es ist für die perorale Gabe geeignet.
	Es hemmt die GTP Synthese und führt so zur Hemmung der DNA-Verlängerung.
	Es hemmt die Protease, die für die Reifung der viralen Proteine essentiell ist.

7	Was sind typische unerwünschte Wirkungen von Cumarinen?	
	×	Harnwegsblutungen
	×	Hirnblutungen
	×	Absinken den Prothrombinspiegels
	×	Teratogenität
	×	Verzögerung der Heilung von Knochenbrüchen

8	Was sind typische unerwünschte Wirkungen von Neuroleptika?	
		Bradykardie
	×	sexuelle Dysfunktion
	×	veränderte endokrine Funktion
	×	Obstipation
	×	orthostatische Hypotension

9	Welche Zuordnungen zwischen Substanz und unerwünschter Wirkung sind zutreffend?	
		Pyrazinamid - periphere Neuritis
	×	Ethambutol - Sehnervschäden
		Isoniazid - Hyperurikämie
		Rifampicin - Ototoxität
		Streptomycin - Induzierung von Cytochrom P450-Enzymen

10	Folgende Applikationsformen vermeiden den First-Pass-Effekt in der Leber. Welche Aussagen sind richtig?	
	×	subkutane Gabe
	×	transdermale Gabe
	×	rektale Gabe
	×	sublinguale Gabe
		orale Gabe

B.4 Zusatzaufgaben 2014

Im Jahr 2014 wurden als Zusatzaufgaben wiederum wie in 2013 nur Aufgaben im *MR*-Format benutzt, wobei allerdings nur die *R4*-Variante zum Einsatz kam. Es kamen, bis auf eine Ausnahme, die gleichen Aufgaben zum Einsatz wie im Vorjahr. Aufgabe 3 mit einer richtigen Alternative wurde durch eine andere Aufgabe mit einer richtigen Alternative ersetzt, da die Formulierung der Aufgabe in 2013 Rückschlüsse darauf zuließ, dass nur eine Alternative korrekt sein kann. Daher ist an dieser Stelle nur die neue Aufgabe 3 wiedergegeben.

3	Welche der genannten Antibiotika wirken durch Hemmung der bakteriellen Proteinbiosynthese?	
		Cefalexin
		Flucloxacillin
		Vancomycin
	×	Clindamycin
		Fosfomycin

C

Begleittexte zu den Zusatzaufgaben

Abgedruckt sind die Begleittexte, welche die Prüflinge auf dem Aufgabenblatt mit den Zusatzfragen in der jeweiligen Prüfung zur Erklärung des Aufgabenformats vorfanden. Hier kursiv hervorgehobene Textstellen waren auch im originalen Begleittext hervorgehoben, 2012 in ***fett-kursiver*** Schreibweise, 2013 und 2014 in *kursiver* Schreibweise.

C.1 Begleittext 2012

Es können jeweils *mehrere Antworten* richtig sein, aber auch *nur eine* oder *keine*.

Bitte lesen Sie die Fragen aufmerksam und antworten Sie, indem Sie für *jede Antwortmöglichkeit einschätzen*, wie sicher Sie sich sind, dass die Antwortmöglichkeit falsch bzw. richtig ist.

C.2 Begleittext 2013

Auf den folgenden Seiten finden Sie zehn allgemeine Prüfungsfragen zur Pharmakologie (Zusatzfragen), einige Fragen zu diesen Zusatzfragen und Fragen zu Ihrem Studierverhalten. Wir möchten Sie bitten, alle diese Fragen gewissenhaft zu beantworten, so dass wir aussagekräftige Ergebnisse zur Verbesserung der Lehre erhalten.

Das Format der zehn Zusatzfragen befindet sich in der Erprobung und weicht von den bisher verwendeten Formaten ab. Daher hat Ihr erzieltes Ergebnis in diesen Zusatzfragen keinen Einfluss auf das Bestehen dieser Klausur. Dennoch ist es von entscheidender Bedeutung, Ihr Klausurergebnis und Ihr Ergebnis bei den Zusatzfragen in Beziehung setzen zu können. *Geben Sie daher bitte in jedem Falle Ihre Matrikelnummer oben an.*

C.2.1 zusätzlich für das *MC*-Format

Die Zusatzfragen sehen auf den ersten Blick aus wie ganz normale Fragen aus der eigentlichen Klausur. Es gibt jedoch einen entscheidenden Unterschied:

Bei den Zusatzfragen ist *nicht* nur genau eine Alternative korrekt, sondern es können mehrere Alternativen korrekt sein. Mindestens eine Alternative ist korrekt, es können aber auch zwei, drei, vier oder alle Alternativen korrekt sein.

Kreuzen Sie daher *alle* Alternativen einer Frage an, welche Sie für eine korrekte Antwort halten und lassen Sie alle Alternativen *frei*, welche Sie für falsch halten!

Für jede richtige Entscheidung (Kreuz bei einer korrekten Alternative bzw. kein Kreuz bei einer falschen Alternative) erhalten Sie einen Punkt.

C.2.2 zusätzlich für das *MTF*-Format

Die Zusatzfragen sehen auf den ersten Blick aus wie ganz normale Fragen aus der eigentlichen Klausur. Es gibt jedoch zwei entscheidende Unterschiede:

1. Bei den Zusatzfragen ist *nicht* nur genau eine Alternative korrekt, sondern es können mehrere Alternativen korrekt sein. Mindestens eine Alternative ist korrekt, es können aber auch zwei, drei, vier oder alle Alternativen korrekt sein.

2. Es gibt für jede Alternative zwei Kästchen, in denen Sie ein Kreuz machen können - eines für *richtig* und eines für *falsch*.

Kreuzen Sie daher bei *allen* Alternativen einer Frage, welche Sie für eine korrekte Antwort halten *richtig* an und kreuzen Sie bei *allen* Alternativen, welche Sie für eine falsche Antwort halten *falsch* an! Achten Sie darauf, dass bei jeder Alternative (also in jeder Zeile) nur ein Kreuz steht. Für jede richtige Entscheidung (*richtig* bei einer korrekten Alternative bzw. *falsch* bei einer falschen Alternative) erhalten Sie einen Punkt.

C.2.3 zusätzlich für das *R4*-Format

Die Zusatzfragen sehen auf den ersten Blick aus wie ganz normale Fragen aus der eigentlichen Klausur. Es gibt jedoch zwei entscheidende Unterschiede:

1. Bei den Zusatzfragen ist *nicht* nur genau eine Alternative korrekt, sondern es können mehrere Alternativen korrekt sein. Mindestens eine Alternative ist korrekt, es können aber auch zwei, drei, vier oder alle Alternativen korrekt sein.

2. Es gibt für jede Alternative vier Kästchen, in denen Sie ein Kreuz machen können - eines für *sicher richtig*, eines für *eher richtig*, eines für *eher falsch* und eines für *sicher falsch*.

Entscheiden Sie daher für *alle* Alternativen einer Frage, ob Sie diese für eine korrekte oder eine falsche Antwort halten. Überlegen Sie dann, ob Sie sich sicher sind in Ihrem Urteil oder eher weniger sicher.

Kreuzen Sie dann bei *allen* Alternativen einer Frage, das entsprechende Kästchen an, also *sicher richtig*, wenn Sie die Alternative für korrekt halten und sich sicher sind, *eher richtig*, wenn Sie die Alternative für korrekt halten und sich eher weniger sicher sind, *eher falsch*, wenn Sie die Alternative für falsch halten und sich eher weniger sicher sind und *sicher falsch*, wenn Sie die Alternative für falsch halten und sich sicher sind.

Achten Sie darauf, dass bei jeder Alternative (also in jeder Zeile) nur ein Kreuz steht. Für jede richtige Entscheidung (*sicher*/*eher richtig* bei einer korrekten Alternative bzw. *sicher*/*eher falsch* bei einer falschen Alternative) erhalten Sie einen Punkt.

C.3 Begleittext 2014

Wie bereits im Aushang zur Klausur bekanntgegeben, finden Sie auf den folgenden Seiten zehn allgemeine Prüfungsfragen zur Pharmakologie (Zusatzfragen) und einige Fragen zu diesen Zusatzfragen. Wir möchten Sie bitten, alle diese Fragen gewissenhaft zu beantworten, so dass wir aussagekräftige Ergebnisse zur Verbesserung der Lehre erhalten. *Durch die korrekte Beantwortung der Zusatzfragen können Sie bis zu zwei Bonuspunkte für die Abschlussklausur erwerben. Geben Sie daher Ihre korrekte Matrikelnummer an, sonst können die Bonuspunkte nicht zugeordnet werden!*

Das Format der Zusatzklausur befindet sich in der Erprobung und weicht von den bisher verwendeten Formaten ab. Die Zusatzfragen sehen auf den ersten Blick aus wie ganz normale Fragen aus der eigentlichen Klausur. Es gibt jedoch zwei entscheidende Unterschiede:

1. Im Gegensatz zu den Ihnen bisher bekannten Fragen ist *nicht* immer genau eine Alternative richtig, sondern es können auch mehrere oder keine der Alternativen richtig sein. Die folgenden Fälle sind möglich: *keine*, genau *eine*, genau *zwei*, genau *drei*, genau *vier* oder *alle fünf* Alternativen sind richtig. Die Anzahl

der richtigen Alternativen einer Frage ist nicht vermerkt. Aus Gründen der Übersichtlichkeit sind alle Fragen im Plural formuliert, auch dann, wenn aus *inhaltlichen Gründen* eine oder keine Alternative korrekt ist.

2. Zur Beantwortung jeder einzelnen Alternative stehen Ihnen vier Kästchen zur Verfügung: eines für *sicher richtig*, eines für *eher richtig*, eines für *eher falsch* und eines für *sicher falsch*. Wählen Sie jeweils das Kästchen aus, welches Ihrer Einschätzung am ehesten entspricht. Machen Sie in jeder Zeile immer nur *genau ein Kreuz*, sonst ist Ihre Antwort automatisch falsch! Lassen Sie auch *keine Zeile aus*, da auch dies automatisch als falsche Antwort gewertet wird.

Alle Alternativen sind immer eindeutig richtig oder eindeutig falsch. Die perfekte Beantwortung der Zusatzfragen würde es daher erfordern, immer sicher richtig bzw. sicher falsch anzukreuzen. Wir möchten Ihnen aber Gelegenheit geben, auszudrücken, ob Sie sich bei Ihrer Antwort sicher sind oder nicht, da dies zusätzliche Informationen über Ihren Wissensstand liefert.

Gehen Sie bei der Beantwortung der Fragen wie folgt vor:

1. Wenn Sie sich sicher sind, dass eine Alternative richtig oder falsch ist, dann kreuzen Sie *sicher richtig* bzw. *sicher falsch* an.

2. Wenn Sie nicht sicher sind, ob eine Alternative richtig oder falsch ist, Sie aber in eine bestimmte Richtung tendieren, kreuzen Sie *eher richtig* bzw. *eher falsch* an.

3. Wenn Sie nicht wissen, ob eine Alternative richtig oder falsch ist, raten Sie und kreuzen Sie *eher richtig* bzw. *eher falsch* an.

Die Bewertung jeder Alternative erfolgt wie folgt:

1. Wenn Sie *sicher richtig* bzw. *sicher falsch* angekreuzt haben und die Alternative ist tatsächlich richtig bzw. falsch, erhalten Sie drei Punkte.

2. Wenn Sie *eher richtig* bzw. *eher falsch* angekreuzt haben und die Alternative ist tatsächlich richtig bzw. falsch, erhalten Sie zwei Punkte.

3. Wenn Sie *eher falsch* bzw. *eher richtig* angekreuzt haben und die Alternative ist tatsächlich richtig bzw. falsch, erhalten Sie dennoch einen Punkt, obwohl Sie falsch geantwortet haben!

4. Wenn Sie *sicher falsch* bzw. *sicher richtig* angekreuzt haben und die Alternative ist tatsächlich richtig bzw. falsch, erhalten Sie keinen Punkt.

D

Aussagen auf dem Evaluationsbogen

Im Folgenden findet sich eine Auflistung aller Aussagen auf dem Evaluationsbogen. Die Aussagen eins bis acht waren für alle Prüflinge gleich und sind in der tatsächlich genutzten Reihenfolge wiedergegeben und zusätzlich entsprechend nummeriert. Die Aussagen waren jeweils mit einem fünfstufigen *Rating* zu beantworten, welches von links nach rechts mit *trifft überhaupt nicht zu, trifft eher nicht zu, weder noch, trifft eher zu* und *trifft völlig zu* betitelt war.

Die Aussagen neun und zehn unterschieden sich für die einzelnen Prüflinge in Abhängigkeit ihrer Gruppenzugehörigkeit, also je nachdem, in welchem *MR*-Antwortformat sie die Zusatzaufgaben erhalten hatten. Die Aussagen waren jeweils so gewählt, dass diese gerade nach den beiden, nicht gerade selbst bearbeiteten Antwortformaten fragten. Die beiden Aussagen waren jeweils mit einem fünfstufigen *Rating* zu beantworten, wobei die Kategorien von links nach rechts mit *viel leichter, etwas leichter, ungefähr gleich, etwas schwerer* und *viel schwerer* betitelt waren.

Die nachstehenden Fragen sollen uns dazu dienen, einzuschätzen, inwieweit das Format der zehn Zusatzfragen, welche Sie gerade beantwortet haben, für den Einsatz in der Lehre und in Prüfungen geeignet ist. Bitte denken Sie bei der Beantwortung der folgenden Fragen an Ihren Eindruck, den Sie beim Beantworten der Zusatzfragen hatten.	
1	Die Beantwortung der Zusatzfragen ist mir insgesamt leicht gefallen.
2	Mit dem Format der Fragen bin ich gut klargekommen.
3	Ich denke, das Format ist generell für den Einsatz in Prüfungen geeignet.
4	Ich würde mir in diesem Fach mehr Prüfungen mit diesem Format wünschen.
5	Ich würde mir auch in anderen Fächern Prüfungen mit diesem Format wünschen.
6	Ich denke, das Format ist generell für den Einsatz in der Lehre (z.B. in ILIAS-Lernmodulen) geeignet.
7	Ich würde mir in diesem Fach mehr Lernmodule mit diesem Format wünschen.
8	Ich würde mir auch in anderen Fächern Lernmodule mit diesem Format wünschen.

Stellen Sie sich nun vor, Sie wären mit einem der unten genannten Formate geprüft worden. Geben Sie jeweils an, wie Ihnen die Beantwortung der Fragen mit dem genannten Format gefallen wäre.	
	Statt wie bisher sollen Sie *alle korrekten* Alternativen ankreuzen und die *falschen* Alternativen *nicht* ankreuzen.
	Statt wie bisher sollen Sie aus zwei Möglichkeiten für jede Alternative wählen: *richtig* oder *falsch*.
	Statt wie bisher sollen Sie aus vier Möglichkeiten für jede Alternative wählen: *sicher richtig*, *eher richtig*, *eher falsch* oder *sicher falsch*.

Abbildungen

1.1 Schematische Darstellung verschiedener Möglichkeiten für Antwort-Wahl-Fragebögen (*MCQ*s) 7

2.1 *Noise*- und Signalverteilung des Signalentdeckungsmodells 28
2.2 *Noise*- und Signalverteilung des Signalentdeckungsmodells mit Wahrscheinlichkeiten 29
2.3 *Noise*- und Signalverteilung des *equal-variance*-Signalentdeckungsmodells 31
2.4 Plot der *hit*- und *false alarm*-Raten aus dem Beispielexperiment nach Wickens (2002) 37
2.5 Plot der *receiver operating characteristic* (*ROC*) aus dem Beispielexperiment nach Wickens (2002) 39
2.6 Plot mehrerer *ROC*-Kurven im *equal-variance*-Modell mit verschiedenen Werten für *d'* 40
2.7 Plot der Iso-Sensitivitätskurven im *equal-variance*- und *unequal-variance*-Modell in Gauss'schen Koordinaten 42
2.8 *Noise*- und Signal-Verteilung beim *Rating*-Verfahren mit vier Antwortkategorien 48
2.9 Plot mehrerer *ROC*-Kurven im *unequal-variance*-Modell mit verschiedenen Werten für μ_s 50

3.1 Beispiel für die Anpassung einer *SR*-Aufgabe an das *MR*-Format . . . 63
3.2 Wahrscheinlichkeitsbaum mit Modellgleichung für p_k nach Lukas (2015a, 2015b) . . . 65
3.3 Beispielaufgabe aus der Prüfung 2012 im *SR*- und *R5*-Format . . . 69
3.4 Gesamtpunkte für die reguläre Prüfung vs. Zusatzaufgaben im Jahr 2012 . . . 74
3.5 Gesamtpunkte für die reguläre Prüfung vs. Signalentdeckungsparameter im Jahr 2012 . . . 75
3.6 Beispielaufgabe aus der Prüfung 2013 im *MC*-, *MTF*- und *R4*-Format . 81
3.7 Gesamtpunkte für die reguläre Prüfung vs. Zusatzaufgaben im Jahr 2013 . . . 87
3.8 Gesamtpunkte für die reguläre Prüfung vs. Signalentdeckungsparameter im Jahr 2013 . . . 89
3.9 Gesamtpunkte für die reguläre Prüfung vs. Zusatzaufgaben im Jahr 2014 . . . 97
3.10 Gesamtpunkte für die reguläre Prüfung vs. Signalentdeckungsparameter im Jahr 2014 . . . 98
3.11 Mittelwertsverläufe der Evaluation im Jahr 2013 und 2014 . . . 102

Tabellen

2.1 Vierfelder-Schema für *hits, misses, false alarms* und *correct rejections* 22
2.2 Daten aus dem Beispielexperiment nach Wickens (2002) 23
2.3 Parameterschätzung aus dem Beispielexperiment nach Wickens (2002) 36
2.4 *Pay-off*-Matrizen für das Beispielexperiment nach Wickens (2002) . . 38
2.5 Beispielhafte Datentabelle für ein *Rating*-Verfahren 47

3.1 Deskriptive Daten der Stichprobe im Jahr 2012 71
3.2 Ergebnisse für die Prüfung und Zusatzaufgaben im Jahr 2012 73
3.3 Deskriptive Daten der Stichprobe im Jahr 2013 83
3.4 Ergebnisse für die Prüfung und Zusatzaufgaben im Jahr 2013 86
3.5 Deskriptive Daten der Stichprobe im Jahr 2014 94
3.6 Ergebnisse für die Prüfung und Zusatzaufgaben im Jahr 2014 96
3.7 Ergebnisse der Evaluation im Jahr 2013 und 2014 101

Literatur

Abdi, H. (2007). Signal detection theory. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 886–889). Thousand Oaks, CA: Sage.

Abele, C., Bargel, H., Pajarinen, A., & Schmidt, M. (2009). *Studienbedingungen und Berufserfolg: Absolventenbefragung der Universität Konstanz - Prüfungsjahrgang 2007*. Konstanz, Germany: Universität Konstanz. Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-103444

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley. doi:10.1002/0471249688

Altman, D. G. (1991). *Practical statistics for medical research*. London, UK: Chapman and Hall.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*(2), 81–99.

Beck, J. R. (1991). Decision-making studies in patient management: Twenty years later. *Medical Decision Making, 11*, 112–115. doi:10.1177/0272989X9101100207

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*(1), 23–35.

Bittrich, K. & Blankenberger, S. (2011). *Experimentelle Psychologie: Ein Methodenkompendium*. Weinheim, Germany: Beltz.

Blackwell, H. R. (1953). *Psychophysical thresholds: Experimental studies of methods of measurement* (Report No. 36). Ann Arbor, MI: University of Michigan, Bulletin of the Engineering Research Institute.

Bland, M. (2000). *An introduction to medical statistics* (3rd ed.). Oxford, UK: Oxford University Press.

Bodde, D. (1948). *Chinese ideas in the west, prepared for the Committee on Asiatic Studies in American Education*. Washington, DC: American Council on Education.

Boneau, C. A. & Cole, J. L. (1967). Decision theory, the pigeon, and the psychophysical function. *Psychological Review, 74*, 123–135. doi:10.1037/h0024287

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin, Germany: Springer. doi:10.1007/978-3-642-12770-0

Broadbent, G. E. & Gregory, M. (1963). Division of attention and the decision theory of signal detection. *Proceedings of the Royal Society of London B: Biological Sciences, 158*, 222–231. doi:10.1098/rspb.1963.0044

Brown, A. S., Schilling, H. E. H., & Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology, 91*, 756–764. doi:10.1037/0022-0663.91.4.756

Bundesministerium der Justiz und für Verbraucherschutz. (2002, June 27). Approbationsordnung für Ärzte vom 27. Juni 2002 (BGBl. I S. 2405), die zuletzt durch Artikel 2 der Verordnung vom 2. August 2013 (BGBl. I S. 3005) geändert worden ist [ÄApprO 2002]. Retrieved from http://www.gesetze-im-internet.de/_appro_2002/BJNR240500002.html

Case, S. M., Becker, D. F., & Swanson, D. B. (1993). Performances of men and women on NBME part I and part II: The more things change... *Academic Medicine, 68*(10 Supplement), S25–7.

Case, S. M. & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Christensen, R. H. B. (2015a). *Analysis of ordinal data with cumulative link models: Estimation with the R-package ordinal*. Retrieved from https://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf

Christensen, R. H. B. (2015b). Ordinal: Regression models for ordinal data (Version 2015.6-28) [Computer Software]. Retrieved from https://cran.r-project.org/web/packages/ordinal/

Christensen, R. H. B. & Brockhoff, P. B. (2015). sensR: Thurstonian models for sensory discrimination (Version 1.4-5) [Computer Software]. Retrieved from https://cran.r-project.org/web/packages/sensR/

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1939). Note on the multiple true-false test exercise. *Journal of Educational Psychology, 30*, 628–631. doi:10.1037/h0058247

Davis, F. B. (1967). A note on the correction for chance success. *The Journal of Experimental Education, 35*, 42–47. doi:10.1080/00220973.1967.11010995

DeCarlo, L. T. (2011). Signal detection theory with item effects. *Journal of Mathematical Psychology, 55*, 229–239. doi:10.1016/j.jmp.2011.01.002

Diedenhofen, B. & Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*, 1–9. doi:10.1027/1015-5759/a000295

Egan, J., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *The Journal of the Acoustical Society of America, 31*, 768–773. doi:10.1121/1.1907783

Electric Paper Evaluationssysteme GmbH. (2015a). EvaExam (Version 6.1) [Computer Software]. Retrieved from http://www.evasys.de

Electric Paper Evaluationssysteme GmbH. (2015b). EvaSys (Version 6.1) [Computer Software]. Retrieved from http://www.evasys.de

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Erickson, M. E. (1972). Test sophistication: An important consideration. *Journal of Reading, 16*(2), 140–144.

Espinosa, M. P. & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*, 415–425. doi:10.1016/j.jmp.2010.06.001

Fechner, G. T. (1860). *Elemente der Psychophysik* (Vols. 2). Leipzig, Germany: Breitenkopf und Härtel.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society, 22*, 700. doi:10.1017/S0305004100009580

Frary, R. B. (1969). Elimination of the guessing component of multiple-choice test scores: Effect on Reliability and Validity. *Educational and Psychological Measurement, 29*, 665–680. doi:10.1177/001316446902900310

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice, 7*, 33–38. doi:10.1111/j.1745-3992.1988.tb00434.x

Friedman, C. (1999). The frequency interpretation in probability. *Advances in Applied Mathematics, 23*, 234–254. doi:10.1006/aama.1999.0653

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 11*, 21–26. doi:10.1111/j.1745-3992.1992.tb00259.x

Galton, F. (1869). *Hereditary Genius*. London, UK: Macmillan and Company.

Goodenough, F. L. (1950). Edward Lee Thorndike: 1874-1949. *The American Journal of Psychology, 63*(2), 291–301.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för Matematik, 1*, 195–277. doi:10.1007/BF02590638

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 37–50. doi:10.1207/s15324818AME0201_3

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333. doi:10.1207/S15324818AME1503_5

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hautus, M. J. (1995). Corrections of extreme proportions and their biasing effects on estimated values of *d′*. *Behavior Research Methods, Instruments, & Computers, 27*(1), 46–51.

Hautus, M. J. & Lee, A. J. (1998). The dispersions of estimates of sensitivity obtained from four psychophysical procedures: Implications for experimental design. *Perception & Psychophysics, 60*(4), 638–649.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Hollander, M. & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York, NY: Wiley.

Hülsheger, U. R., Maier, G. W., & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment, 15*, 3–18. doi:10.1111/j.1468-2389.2007.00363.x

Hutchinson, T. P. (1981). A review of some unusual applications of signal detection theory. *Quality and Quantity, 15*, 71–98. doi:10.1007/BF00144302

ILIAS open source e-Learning e.V. (2015). ILIAS eA (Version 5.0) [Computer software]. Retrieved June 24, 2015, from http://www.ilias.de

Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP). (2015). *Gegenstandskataloge*. Retrieved from http://www.impp.de/

Jones, F. N. (1956). A forced-choice method of limits. *The American Journal of Psychology*, *69*, 672. doi:10.2307/1419098

Kampmeyer, D., Matthes, J., & Herzig, S. (2014). Lucky guess or knowledge: A cross-sectional study using the Bland and Altman analysis to compare confidence-based testing of pharmacological knowledge in 3rd and 5th year medical students. *Advances in Health Sciences Education: Theory and Practice*. doi:10.1007/s10459-014-9537-1

Kaplan, R. M. & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth Cengage Learning.

Katkov, M., Tsodyks, M., & Sagi, D. (2006). Analysis of a two-alternative force-choice signal detection theory model. *Journal of Mathematical Psychology*, *50*, 411–420. doi:10.1016/j.jmp.2005.11.002

Kendall, M. G. (1949). On the reconciliation of theories of probability. *Biometrika*, *36*, 101. doi:10.2307/2332534

Klein, S. A. & Macmillan, N. A. (Eds.). (2001). Psychometric functions and adaptive methods [Special Issue]. *Perception & Psychophysics*, *63*(8).

Kramer, J. (2009). Allgemeine Intelligenz und beruflicher Erfolg in Deutschland. *Psychologische Rundschau*, *60*, 82–98. doi:10.1026/0033-3042.60.2.82

Krebs, R. (2004). *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung*. Universität Bern. Retrieved from http://www.iml.unibe.ch/dienstleistung/assessment_pruefungen/pruefungsmethoden/wahlantwortfragen_mc/

Krohne, H. W. & Hock, M. (2007). *Psychologische Diagnostik: Grundlagen und Anwendungsfelder*. Stuttgart, Germany: Kohlhammer.

Krüger, M. (2013). *elsa Handreichung zum Erstellen und Bewerten von Multiple-Choice-Aufgaben*. Leibniz Universität Hannover. Retrieved from https://www.uni-hannover.de/fileadmin/luh/content/elearning/practicalguides2/didaktik/elsa_handreichung_zum_erstellen_und_bewerten_von_mc-fragen_2013.pdf

Kubinger, K. D. (2014). Gutachten zur Erstellung gerichtsfester Multiple-Choice-Prüfungsaufgaben. *Psychologische Rundschau*, *65*, 169–178. doi:10.1026/0033-3042/a000218

Kubinger, K. D. & Gottschall, C. H. (2007). Item difficulty of multple choice tests dependant on different item response formats: An experiment in fundamental research on psychological assessment. *Psychology Science, 49*(4), 361–374.

Lalkhen, A. G. & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain, 8*, 221–223. doi:10.1093/bjaceaccp/mkn041

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6th ed.). Weinheim, Germany: Beltz.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*, 7–11. doi:10.1111/j.1745-3984.1975.tb01003.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lukas, J. (2006). Signalentdeckungstheorie. In J. Funke & P. Frensch (Eds.), *Handbuch der Psychologie: Vol. 5: Handbuch der Allgemeinen Psychologie - Kognition* (pp. 732–740). Göttingen, Germany: Hogrefe.

Lukas, J. (2015a). *Auswertung von Multiple-Choice-Klausuren: Maluspunkte und Ratewahrscheinlichkeiten aus kognitionspsychologischer Perspektive*. Lernen - Verstehen - Wissen: Zweites Wissenschaftliches Kolloquium. Halle, Germany. Retrieved from http://www.llz.uni-halle.de/veranstaltungen/rueckblick/mlg15/kolloquium15/programm/lukas/

Lukas, J. (2015b). *Ratewahrscheinlichkeiten und Maluspunkte bei Multiple-Choice-Aufgaben: Warum gibt es dazu so viele Fehlkonzepte? Und wie macht man es (begründbar) richtig?* e-Prüfungs-Symposium ePS. Paderborn, Germany. Retrieved from http://www.e-pruefungs-symposium.de/wp-content/uploads/2015/11/Abstractband-ePS2015.pdf

Lusted, L. B. (1971a). Decision-making studies in patient management. *The New England Journal of Medicine, 284*, 416–424. doi:10.1056/NEJM197102252840805

Lusted, L. B. (1971b). Signal detectability and medical decision-making. *Science, 171*, 1217–1219. doi:10.1126/science.171.3977.1217

Lyerly, S. B. (1951). A note on correcting for chance success in objective tests. *Psychometrika, 16*, 21–30. doi:10.1007/BF02313424

Macher, S. (2005). *Standardisierte Prüfungsmethoden in der medizinischen Ausbildung: Kapitel IV Multiple Choice-Aufgaben*. Medizinische Universität Graz. Retrieved from https://www.medunigraz.at/fileadmin/lehren/planen-organisieren/pdf/QM_SM_HandbuchPruefungsmethoden_20050404_01.pdf

Macmillan, N. A. & Creelman, C. D. (1991). *Detection theory: A users guide*. New York, NY: Cambridge University Press.

Macmillan, N. A. & Creelman, C. D. (2010). *Detection theory: A user's guide* (2nd ed.). New York, NY: Psychology Press.

Madaus, G. F. & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688–695.

Marcum, J. I. (1947). *A statistical theory of target detection by pulsed radar*. Santa Monica, CA: RAND Corporation.

Marill, T. M. (1956). *Detection theory and psychophysics* (Technical Report No. 319). Boston, MA: Massachusetts Institute of Technology, Research Laboratories of Electronics.

Martin-Luther-Universität Halle-Wittenberg. (2009). Studienordnung für den Studiengang Medizin an der Martin-Luther-Universität Halle-Wittenberg. *Uni-Amtsblatt, 19*(8), 1–22. Retrieved from http://www.verwaltung.uni-halle.de/KANZLER/ZGST/ABL/2009/09_08_01.pdf

Martin-Luther-Universität Halle-Wittenberg. (2012). Zweite Ordnung zur Änderung der Studienordnung für den Studiengang Medizin an der Martin-Luther-Universität Halle-Wittenberg. *Uni-Amtsblatt, 22*(11), 1–6. Retrieved from http://www.verwaltung.uni-halle.de/KANZLER/ZGST/ABL/2012/12_11_02.pdf

Mathews, J. (2006, November 14). Just whose idea was all this testing? *The Washington Post*. Retrieved from http://www.washingtonpost.com

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McKenzie, C. R., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes-no and forced-choice tasks. *Journal of Experimental Psychology: General, 130*(1), 140–155.

McNicol, D. (2005). *A primer of signal detection theory*. Mahwah, NJ: Lawrence Erlbaum Associates.

McPhail, I. P. (1979). Test sophistication: An important consideration in judging the standardized test performance of black students. *Reading World, 18*, 227–235. doi:10.1080/19388077909557479

Meter, D. van & Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Transactions of the IRE Professional Group on Information Theory, 4*, 119–145. doi:10.1109/TIT.1954.1057471

Miles, J. (1973). Eliminating the guessing factor in the multiple choice test. *Educational and Psychological Measurement, 33*, 637–651. doi:10.1177/001316447303300313

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

Mises, R. von. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Wien, Austria: Springer.

Much, S. (2014). *Die Eignung probabilistischer Testmodelle zur Wissensdiagnostik mit Multiple-Choice-Klausuren: Polytome Item-Response-Modelle und Signal-Entdeckungs-Theorie.* Unpublished master's thesis, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany.

Munson, W. A. & Karlin, J. E. (1954). The measurement of human channel transmission characteristics. *The Journal of the Acoustical Society of America, 26*, 542–553. doi:10.1121/1.1907372

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 236*, 333–380. doi:10.1098/rsta.1937.0005

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289–337.

Organisation for Economic Cooperation and Development. (2012). *PISA 2012 technical report*. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Cooperation and Development. (2015). *Programme for International Student Assessment (PISA)*. Retrieved from http://www.oecd.org/pisa/

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Evaluation in education and human services. Boston, MA: Kluwer Academic Publishers.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50.

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review*, *73*(1), 44–58.

Paulsen, F. (1902). *Die deutschen Universitäten und das Universitätsstudium*. Berlin, Germany: A. Asher & Co.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Proceedings of the IRE Professional Group on Information Theory*, *4*, 171–212. doi:10.1109/TIT.1954.1057460

Plumlee, L. B. (1952). The effect of difficulty and chance success on item-test correlation and on test reliability. *Psychometrika*, *17*, 69–86. doi:10.1007/BF02288796

Plumlee, L. B. (1954). The predicted and observed effect of chance success on multiple-choice test validity. *Psychometrika*, *19*, 65–70. doi:10.1007/BF02288994

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: Précédées des règles générales du calcul des probabilités*. Paris, France: Bachelier.

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory*. Cambridge, MA: MIT Press.

Preacher, K. J. (2002). Calculation for the test of the difference between two independent correlation coefficients [Computer software]. Retrieved from http://quantpsy.org/

Preston, R. C. (1965). The multiple-choice test as an instrument in perpetuating false concepts. *Educational and Psychological Measurement*, *25*(1), 111–116.

R Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.2) [Computer Software]. Retrieved from http://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago, IL: University of Chicago Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*, 3–13. doi:10.1111/j.1745-3992.2005.00006.x

Roediger, H. L., III & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159. doi:10.1037/0278-7393.31.5.1155

Rogers, W. T. & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*, 247–259. doi:10.1027/1015-5759.12.3.247

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2nd ed.). Bern, Switzerland: Huber.

Rowley, G. L. (1974). Which examinees are most favored by the use of multiple choice tests? *Journal of Educational Measurement*, *11*(1), 15–23.

Runté, R. (2001). Basic dos and don'ts of multiple-choice examinations. St. John's, NL, Canada.

Sacks, H. S., Chalmers, T. C., & Smith, H., Jr. (1983). Sensitivity and specificity of clinical trials. *Archives of Internal Medicine*, *143*, 753. doi:10.1001/archinte.1983.00350040143020

Schmidts, M. & Lischka, M. (2001). *Prüfungsfragen für Multiple-Choice Tests erstellen*. Universität Wien. Retrieved from http://www.med.uni-giessen.de/intranet/lehre/Anleitung_Erstellung_von_MC-Fragen.pdf

Seneta, E. (2013). A tricentenary history of the law of large numbers. *Bernoulli, 19*, 1088–1121. doi:10.3150/12-BEJSP12

Smith, M. & Wilson, E. A. (1953). A model of the auditory threshold and its application to the problem of the multiple observer. *Psychological Monographs: General and Applied, 67*, 1–35. doi:10.1037/h0093654

Spencer, H. (1855). *Principles of psychology*. London, UK: Longman, Brown, Green, and Longmans.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE Life Sciences Education, 11*, 294–306. doi:10.1187/cbe.11-11-0100

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149.

Swets, J. A. (1961). Is there a sensory threshold? *Science, 134*, 168–177. doi:10.1126/science.134.3473.168

Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers: Contemporary readings*. New York, NY: Wiley.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Lawrence Erlbaum Associates.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*, 301–340. doi:10.1037/h0040547

Tanner, W. P., Jr. & Swets, J. A. (1953). *A new theory of visual detection* (Technical Report No. 18). Ann Arbor, MI: University of Michigan, Electronic Defense Group.

Tanner, W. P., Jr. & Swets, J. A. (1954a). A decision-making theory of visual detection. *Psychological Review, 61*, 401–409. doi:10.1037/h0058700

Tanner, W. P., Jr. & Swets, J. A. (1954b). The human use of information I: Signal detection for the case of the signal known exactly. *Transactions of the IRE Professional Group on Information Theory, 4*, 213–221. doi:10.1109/TIT.1954.1057461

Têng, S.-y. (1943). Chinese influence on the western examination system: I. Introduction. *Harvard Journal of Asiatic Studies, 7*, 267–312. doi:10.2307/2717830

The Japan Foundation and Japan Educational Exchanges and Services. (2011). *Scaled scores*. Retrieved from http://www.jlpt.jp/e/about/pdf/scaledscore_e.pdf

The Japan Foundation and Japan Educational Exchanges and Services. (2015). *Official worldwide Japanese-Language Proficiency Test website*. Retrieved from http://www.jlpt.jp

Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review, 34*, 273–286. doi:10.1037/0033-295X.101.2.266

Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology, 38*, 368–389.

Toppino, T. C. & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research, 86*, 357–362. doi:10.1080/00220671.1993.9941229

Towns, M. H. & Robinson, W. R. (1993). Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching, 30*, 709–722. doi:10.1002/tea.3660300709

Treisman, A. & Geffen, G. (1967). Selective attention: Perception or response? *Quarterly Journal of Experimental Psychology, 19*, 1–18. doi:10.1080/14640746708400062

Treisman, M. & Watts, T. R. (1966). Relation between signal detectability theory and the traditional procedures for measuring sensory thresholds: Estimating d′ from results given by the method of constant simuli. *Psychological Bulletin, 66*, 438–454. doi:10.1037/h0020413

Tsirelson, B. (2012, May 13). Law of large numbers. *Encyclopedia of Mathematics*. Retrieved from http://www.encyclopediaofmath.org/index.php?title=Law_of_large_numbers&oldid=26552

Tyler, C. W. & Chen, C.-C. (2000). Signal detection theory in the 2AFC paradigm: Attention, channel uncertainty and probability summation. *Vision Research, 40*, 3121–3144. doi:10.1016/S0042-6989(00)00157-7

Vernon, P. E. (1938). Intelligence test sophistication. *British Journal of Educational Psychology, 8*, 237–244. doi:10.1111/j.2044-8279.1938.tb03129.x

Vernon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22*, 269–286. doi:10.1177/001316446202200203

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics, 10*, 299–326. doi:10.1214/aoms/1177732144

Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.

Wald, A. (1950). *Statistical decision functions*. New York, NY: Wiley.

Wang, R. (2013). *The Chinese imperial examination system: An annotated bibliography*. Lanham, MD: Scarecrow Press.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford, UK: Oxford University Press.

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education: Theory and Practice, 20*, 247–263. doi:10.1007/s10459-014-9528-2

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176. doi:10.1037/0033-295X.114.1.152

Yoakum, C. S. & Yerkes, R. M. (1920). *Army mental tests*. New York, NY: Henry Holt and Company.

Zazzo, R. (1993). Alfred Binet. *The Quarterly Review of Comparative Education, 23*, 101–112.

Zimmerman, D. W. & Williams, R. H. (1965). Chance success due to guessing and nonindependence of true scores and error scores in multiple-choice tests: computer trials with prepared distributions. *Psychological Reports, 17*, 159–165. doi:10.2466/pr0.1965.17.1.159

Zimmerman, D. W. & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*, 357–371. doi:10.1177/0146621603254799