AUSWERTUNG VON KLAUSUREN IM ANTWORT-WAHL-FORMAT

J. LUKAS, A. MELZER & S. MUCH

unter Mitarbeit von S. Eisentraut

LLZ
ZENTRUM FÜR MULTIMEDIALES LEHREN UND LERNEN

MARTIN-LUTHER UNIVERSITÄT HALLE-WITTENBERG

Josef Lukas, Andreas Melzer & Sören Much
unter Mitarbeit von Sarah Eisentraut

Auswertung von Klausuren im Antwort-Wahl-Format

Autoren
Prof. Dr. Josef Lukas
Dr. Andreas Melzer
Sören Much, M.Sc.
unter Mitarbeit von Sarah Eisentraut, M.Sc.

Kontakt
Prof. Dr. Josef Lukas
josef.lukas@psych.uni-halle.de

Herausgegeber

@LLZ | Zentrum für multimediales Lehren und Lernen
Hoher Weg 8
06120 Halle (Saale)
+49 (0)345 - 55 28671

Das @LLZ | Zentrum für multimediales Lehren und Lernen ist eine zentrale Einrichtung der
Martin-Luther-Universität Halle-Wittenberg
Körperschaft des öffentlichen Rechts
vertreten durch den Rektor Prof. Dr. Udo Sträter
Universitätsplatz 10
06108 Halle (Saale)

Bundesministerium für Bildung und Forschung

Gemeinsames Bund-Länder-Programm für bessere Studienbedingungen und mehr Qualität in der Lehre.
Dieses Vorhaben wird aus Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01PL17065 gefördert.
Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Coverdesign: Melanie Grießer
Korrektur: Alfred Kuhn, M.A.
URN: urn:nbn:de:gbv:3:2-66099
ISBN: 978-3-86829-873-4
Diese elektronische Publikation ist auch als Buch mit der ISBN 978-3-86829-872-7 erhältlich.
Zitationshinweis: Lukas J., Melzer, A. & Much, S., unter Mitarbeit von S. Eisentraut (2017). *Auswertung von Klausuren im Antwort-Wahl-Format*. Zugriff unter http://nbn-resolving.de/urn:nbn:de:gbv:3:2-66099

Vorwort

Klausuren mit Aufgaben nach dem „Antwort-Wahl-Verfahren" („AWV"), wie die offizielle juristische Bezeichnung für Aufgaben lautet, bei denen eine oder mehrere richtige Antworten aus einer Reihe von vorgegebenen Antwortmöglichkeiten auszuwählen sind, nehmen immer breiteren Raum im akademischen Prüfungsgeschehen ein. Die Gründe dafür sind vielfältig:

- Hohe Studierendenzahlen bei einer gleichzeitigen Vervielfältigung der Anzahl von Prüfungen im Verlauf eines Studiums erfordern ökonomische Prüfungsverfahren. Die Möglichkeit zur automatischen Auswertung verspricht zumindest für die Bewertung einer Prüfung eine drastische Reduzierung des Prüfungsaufwandes.
- Die Auswertung einer Antwort-Wahl-Klausur geht nicht nur schnell, sondern sie ist auch objektiv, d. h., das Ergebnis ist nicht abhängig von der Person des Auswertenden.
- Das gesamte Prüfungsgeschehen ist transparenter als bei herkömmlichen Verfahren der Beurteilung einer Prüfungsleistung durch Prüfende und lässt sich damit viel einfacher in seiner Qualität und Angemessenheit beurteilen.
- Neue Medien und Technologien erlauben auch inhaltlich anspruchsvolle neue Prüfungsformen.

Die große Herausforderung bei der Gestaltung von Antwort-Wahl-Klausuren liegt in der Formulierung von geeigneten Prüfungsaufgaben. Was aber sind „gute Prüfungsaufgaben"? Welche Arten von Wissen, Fertigkeiten, Kompetenzen etc. lassen sich mit Aufgaben im Antwort-Wahl-Format überhaupt erfassen?

Das hier vorgelegte Handbuch zur Prüfungsauswertung wird diese „großen Fragen" nicht oder nur am Rande beantworten. Das Ziel ist zunächst sehr viel bescheidener. Wir nähern uns dem Problem vom Ende des Prüfungsprozesses und konzentrieren uns auf die Frage: Wie werden Antwort-Wahl-Klausuren am besten ausgewertet, d. h., wie lassen sich theoretisch begründbar Punkte und Noten bei Antwort-Wahl-Klausuren vergeben? Das klingt unspekta-

kulär, aber schon ein kurzer Blick in die vielen Handbücher und Ratgeber für die Gestaltung von Antwort-Wahl-Klausuren (z. B. Brüstle, 2011; Case & Swanson, 2002; Haladyna, Downing & Rodriguez, 2002; Haladyna & Rodriguez, 2013), in Prüfungsordnungen deutscher Hochschulen oder in Wikis einschlägiger Internetportale (z. B. ELAN e.V., 2016) zeigt: die Frage der sachgerechten Auswertung von Antwort-Wahl-Klausuren ist geprägt von Unsicherheit und manifesten Fehlkonzepten. Der – im Übrigen sehr verdienstvolle und lesenswerte – Überblicksartikel von Lindner, Strobel und Köller (2015) oder das Gutachten von Kubinger (2014) aus der Perspektive der *Item-Response*-Theorie haben daran bislang wenig geändert. Dabei kann man gerade in diesem Bereich klare und begründbare Antworten geben und allen Prüfenden eine pragmatische Orientierung verschaffen.

Im Zentrum unserer Überlegungen steht die Frage, wie man mit dem Hauptproblem von Aufgaben im Antwort-Wahl-Format umgeht, der typischerweise hohen Ratewahrscheinlichkeit. Darunter versteht man ein bekanntes und leicht verständliches Problem: Auch wer die Antwort bei einer Aufgabe nicht weiß, hat eine mehr oder weniger große Chance, die richtige Antwort zufällig auszuwählen. Prüfende können umgekehrt einer richtigen Antwort nicht ansehen, ob sie gewusst oder nur geraten wurde. Das ist unschön und störend für den eigentlichen Zweck einer Prüfung, nämlich den tatsächlichen Wissenszustand eines Prüflings festzustellen. Die Ratewahrscheinlichkeit gilt deshalb als „Übel", dem die Mehrzahl der Handbücher und Ratgeber mit dem Versuch beikommen will, sie zu minimieren, z. B. durch

- die Gestaltung der Aufgaben (mehr Distraktoren, mehr Antwortoptionen etc.),
- den Verzicht auf Aufgabentypen mit hoher Ratewahrscheinlichkeit,
- das „Bestrafen" von Ratestrategien (z. B. durch sogenannte „Maluspunkte") oder
- spezielle Verfahren der Punktvergabe (Punkte nur für komplett richtige Antwortmuster etc.).

Bei diesem „Kampf gegen das Übel der Ratewahrscheinlichkeit" wird oft übersehen, dass die Reduzierung der Ratewahrscheinlichkeit regelmäßig durch erhebliche Nachteile erkauft wird:

- Höhere Hürden für die Vergabe von Punkten, z. B. wenn es Punkte nur für fehlerfreie Komplettlösungen gibt, reduzieren zwar die Ratewahrscheinlichkeit, sie erhöhen aber gleichzeitig die Wahrscheinlichkeit für den „Fehler zweiter Art". Gemeint ist damit die Wahrscheinlichkeit dafür, dass Prüflinge, die eigentlich über genügend Wissen verfügen, dennoch keine Punkte erhalten, weil sie z. B. in der Zeile verrutschen, etwas falsch verstehen, zu kompliziert denken etc.

- Die Warnung vor Aufgabentypen mit hoher Ratewahrscheinlichkeit veranlasst viele Prüfende dazu, auf Aufgaben, die inhaltlich angemessen wären, z. B. einfache „Ja/Nein"- oder „richtig/falsch"-Aufgaben, zu verzichten. Umgekehrt führt die strikte Orientierung am Kriterium einer niedrigen Ratewahrscheinlichkeit häufig zu inhaltlich fragwürdigen und gelegentlich absurden Aufgabenformulierungen.
- Maßnahmen zur Entmutigung von Ratestrategien sind problematisch, weil oft unklar ist, wie Prüflinge damit umgehen.

Wir werden in diesem Handbuch die Begründungen für diese Einschätzung im einzelnen darlegen und auf dieser Grundlage eine gänzlich andere Strategie empfehlen: Ratewahrscheinlichkeiten sind ein unvermeidbarer Bestandteil von Antwort-Wahl-Klausuren. Sie müssen deshalb bei der Bewertung der Prüfungsleistung berücksichtigt werden. Die Ratewahrscheinlichkeit wird nicht bekämpft, sondern möglichst präzise definiert und bei der Bewertung der Prüfungsleistung in Rechnung gestellt. Die Methoden dazu sind seit langem bekannt und fester Bestandteil der psychologischen Testtheorie (Lord & Novick, 1968) und der Wahrscheinlichkeitstheorie.

Der Vorteil dieses Vorgehens liegt auf der Hand:

- Es gibt keine „verbotenen" Aufgabentypen mehr. Die Formulierung von Aufgaben orientiert sich ausschließlich an den inhaltlichen Anforderungen. Die Höhe der Ratewahrscheinlichkeit ist kein Gütekriterium mehr.
- Raten als Strategie wird nicht mehr entmutigt, sondern im Gegenteil ausdrücklich empfohlen. In der experimentellen Psychologie ist dies aus guten Gründen seit jeher Bestandteil jeder Versuchsanweisung: „Gib Deine Antwort nach bestem Wissen und Gewissen. Wenn Du unsicher bist, wähle die Antwort, die am ehesten in Frage kommt. Wenn Du überhaupt nichts weißt, rate". Dadurch werden vor allem gleiche Verhältnisse für alle Teilnehmer an einer Prüfung geschaffen.
- Die Ratewahrscheinlichkeit wird nicht durch inhaltlich fragwürdige und juristisch problematische (OVG Nordrhein-Westfalen, 2008, Rn. 70 & 74; VG Arnsberg, 2012, Rn. 116ff) Maluspunkte kompensiert, sondern wird bei der Festsetzung der Bestehens- und Notengrenzen berücksichtigt.
- Auch wenn die Ratewahrscheinlichkeit unterschiedlicher Aufgaben innerhalb einer Klausur durch die Verwendung von verschiedenen Aufgabentypen variiert oder gar auf null absinkt, lässt sich ein einheitliches Bewertungsverfahren verwenden.

Das Handbuch ist in drei Teilen organisiert: In Teil I werden die theoretischen Grundlagen für die hier skizzierte Argumentation besprochen. Zunächst wird ein einfaches, probabilistisches Modell für das Lösen von Aufgaben formuliert. Dieses Modell liefert mit der strengen begrifflichen Unterscheidung von „wissen“ und „richtig antworten“ die Grundlage für eine am Wissen orientierte Auswertung von Aufgaben im Allgemeinen und Antwort-Wahl-Formaten im Speziellen. Anschließend werden die verschiedenen, gängigen Aufgabenformate und deren Unterscheidungsmerkmale dargestellt und eine Zuordnung der ILIAS-Aufgabentypen (ILIAS open source e-Learning e.V., 2017) zu den Aufgabenformaten vorgenommen, die abschließend in Tabelle 5.1 übersichtlich zusammengestellt werden.

In Teil II werden die im ersten Teil entwickelten theoretischen Grundlagen auf die verschiedenen Aufgabenformate angewandt. Die Aufgabenformate werden im Detail beschrieben, es werden Vor- und Nachteile diskutiert und die Anwendung der Auswertungsprinzipien aus Teil I wird an konkreten Beispielen erläutert. Daraus werden Empfehlungen für die Verwendung abgeleitet.

Im letzten Teil III sind die im zweiten Teil gewonnenen Erkenntnisse über die Auswertung der einzelnen Aufgabenformate in einer Art Kurzreferenz, sogenannten *Cheat Sheets*, mit allen nötigen Informationen für die schnelle Verwendung im Prüfungsalltag zusammengefasst. Wer sich weniger für die Begründung des Verfahrens interessiert und es lediglich anwenden möchte, findet dafür in den *Cheat Sheets* die entsprechenden Handlungsanweisungen. Darüber hinaus ist auf der Internetpräsenz des @LLZ ein Berechnungstool mit den wichtigsten Formeln verfügbar. Werden in dieses Tool die einzelnen Aufgaben mit ihren Parametern und ein Notenschlüssel eingetragen, erfolgt automatisch die Berechnung der Notengrenzen.

Halle (Saale), 10. Oktober 2017

Inhalt

Teil I

Theoretischer Hintergrund

1 Überblick

Prüfungen im Allgemeinen und AW-Klausuren im Besonderen haben das Ziel, das „Wissen" oder die Fertigkeiten von Studierenden festzustellen und in Bewertungskategorien abzubilden. Bei den Bewertungskategorien handelt es sich häufig um Noten zwischen 1.0 („sehr gut") und 5.0 („nicht bestanden") oder auch nur um die beiden Kategorien „bestanden" und „nicht bestanden". Im ersten Teil dieses Handbuchs soll dieser Prozess vom „Wissen" zur „Bewertung" aus kognitionspsychologischer und testtheoretischer Perspektive betrachtet werden, um daraus begründbare Aussagen für eine sachgerechte Auswertung von AW-Klausuren ableiten zu können.

Der Gesamtprozess lässt sich in vier Phasen gliedern (s. Abbildung 1.1). Jedem Übergang von einer Phase zur nächsten ist ein eigenes Kapitel gewidmet:

1: Vom „Wissen" zur Prüfungsleistung. Was und wie viel Prüflinge tatsächlich wissen oder können, ist im Allgemeinen nicht direkt beobachtbar. Beobachtbar ist immer nur das

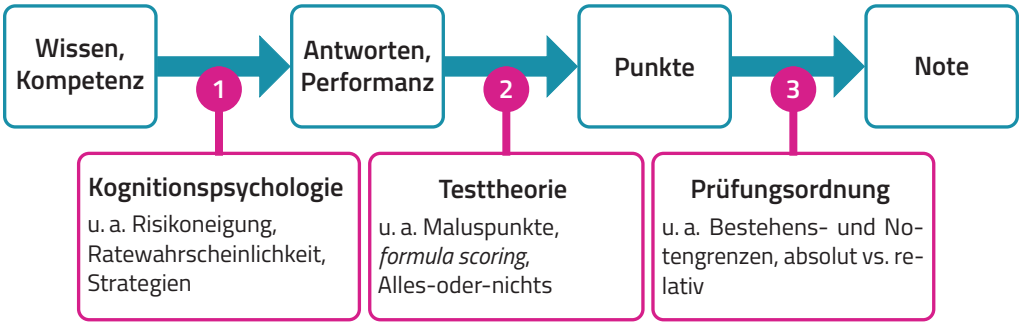


Abbildung 1.1. Vom Wissen zur Note: Übersicht über die vier Phasen des Benotungsprozesses beim Prüfen und die damit verbundenen Fragestellungen aus der Kognitionspsychologie und der psychologischen Testtheorie.

manifeste Verhalten (z. B. in einer Prüfungssituation). Beide Begriffe müssen streng unterschieden werden: Auf der einen Seite das tatsächliche Wissen, die Kompetenz, die die eigentlich interessierende Variable darstellt, aber nicht direkt beobachtbar ist. Auf der anderen Seite die Performanz in der Prüfungssituation, also die Prüfungsleistung. Schon der Begriff der „Ratewahrscheinlichkeit", der bei AW-Klausuren eine prominente Rolle spielt, setzt diese Unterscheidung implizit voraus. In Kapitel 2 wird ein einfaches wahrscheinlichkeitstheoretisches Modell für den Zusammenhang von „Wissen" und „Antwortverhalten" formuliert, das die Grundlage für die Definition von Ratewahrscheinlichkeiten und für einen rational begründbaren Umgang damit bildet.

2: Scoring: Vergabe von Punkten für Antworten. Wie viele Punkte werden für richtige bzw. falsche Antworten vergeben? Soll es Punktabzug, sogenannte „Maluspunkte", für falsche Antworten geben? Und wie verfährt man bei nicht beantworteten Aufgaben? Die Frage nach dem korrekten Scoring, also der Regel, nach der Punkte für die verschiedenen Antwortmöglichkeiten vergeben werden, ist in der psychologischen Testtheorie kontrovers diskutiert worden (Lesage, Valcke & Sabbe, 2013; Lindner et al., 2015; Lord, 1975) und wird im Zusammenhang mit der Auswertung von AW-Klausuren bis heute aus verschiedenen Perspektiven unterschiedlich beantwortet. Rationale Argumentation ist dabei eher die Ausnahme als die Regel (Bar-Hillel, Budescu & Attali, 2005). In Kapitel 3 werden die wichtigsten Scoringverfahren vorgestellt und gezeigt, welchen Einfluss sie auf den Zusammenhang zwischen „Wissen" und der zu erwartenden Punktzahl für das Klausurergebnis haben. Daraus lassen sich objektivierbare Kriterien für die Eignung der verschiedenen Scoringverfahren für unterschiedliche Klausursituationen gewinnen.

3: Bestehensgrenzen und Benotung. Wie viele Punkte sind zum Bestehen einer Klausur notwendig und wie werden die Punkte in Noten umgerechnet? Auf der Grundlage des Modells aus Kapitel 2 und der Scoringverfahren aus Kapitel 3 wird in Kapitel 4 ein einfaches Verfahren zur Bestimmung der Bestehens- und Notengrenzen beschrieben, das die Ratewahrscheinlichkeiten berücksichtigt, (vermutlich) gerichtsfest ist und an die jeweiligen Bestimmungen der Prüfungsordnung angepasst werden kann.

Zum Abschluss des theoretischen Teils gibt Kapitel 5 einen Überblick über die wichtigsten Aufgabenformate für AW-Klausuren, wie sie zum Beispiel in der E-Learning-Plattform ILIAS zur Verfügung gestellt werden (ILIAS open source e-Learning e.V., 2017).

2 Ratewahrscheinlichkeit

Zusammenhang von „wissen“ und „richtig antworten“

Das offensichtlichste und bekannteste Problem bei AW-Klausuren ist die hohe Ratewahrscheinlichkeit: Gemeint ist damit, dass man wegen der begrenzten Anzahl an vorgegebenen Antwortalternativen die richtige auch durch „zufälliges Raten“, also ohne jedes Wissen auswählen kann. Einer richtigen Antwort ist aber im Allgemeinen nicht anzusehen, ob sie auf „echtem Wissen“ oder auf „zufälligem Raten“ beruht.

Die Mehrzahl von Prüfenden (und Prüfungsratgebern) fassen deshalb Raten als lästiges, den Diagnoseprozess störendes Übel auf und versuchen, es mit mehr oder weniger tauglichen Mitteln zu bekämpfen: Raten wird bestraft (z. B. durch Maluspunkte für falsche Antworten), korrigiert (ebenfalls durch Maluspunkte), erschwert (durch Erhöhung der Anzahl an vorgegebenen Alternativen) oder in seiner Wirkung reduziert (durch erhöhte Anforderungen an die Vergabe von Punkten). Vieles davon hat mit der Vergabe von Punkten für richtige bzw. falsche Antworten zu tun, dem sogenannten Scoring. Bevor wir das systematisch im nächsten Kapitel untersuchen (können), müssen wir genauer klären, was unter „wissen“ und „richtig antworten“ zu verstehen ist, wie beides zusammenhängt und was genau eigentlich unter einer „Ratewahrscheinlichkeit“ zu verstehen ist.

2.1 Kompetenz vs. Performanz

Die Unterscheidung von „wissen“ und „richtig antworten“ entspricht in der empirischen Psychologie (wie in jeder empirischen Wissenschaft) der Unterscheidung zwischen einer theoretischen, hypothetischen, nicht direkt beobachtbaren Variablen (z. B. Intelligenz, Temperatur, Wahrscheinlichkeit) und einem empirischen, direkt beobachtbaren Sachverhalt (Anzahl richtiger Antworten in einem Intelligenztest, Ausdehnung einer Quecksilbersäule, relative Häufigkeit). In der auf Chomsky (1965) zurückgehenden Begrifflichkeit entspricht das „Wissen“

der „Kompetenz“ von Prüflingen, die wohl zu unterscheiden ist von deren „Performanz“ in der Prüfungssituation, also dem beobachtbaren Verhalten, den manifesten Auswahlkreuzen auf dem Klausurbogen.

Für ein möglichst einfaches Modell zum Zusammenhang von Kompetenz und Performanz in diesem Sinn unterscheiden wir – angelehnt an das klassische Alles-oder-nichts-Modell von Bower (1961) – für jede Aufgabe nur zwei Wissenszustände für die Kompetenz, nämlich:

- „Wissen“ (W) und
- „Nichtwissen“ (NW)

sowie drei mögliche Antwortkategorien für die Performanz:

- Antwort richtig (A_r),
- Antwort falsch (A_f) und
- keine Antwort (A_o).

2.2 Ein einfaches probabilistisches Modell

Um den (Wissens-)Zuständen und den Antworten Wahrscheinlichkeiten zuordnen zu können, definieren wir ein Zufallsexperiment, das etwa so aussehen könnte: Für eine Klausur (z. B. zum Modul „Allgemeine Psychologie II“) denken wir uns eine Menge von Klausuraufgaben, die den Stoff definieren. Aus dieser (potenziell unendlichen) Menge wird zufällig eine Aufgabe ausgewählt und den Prüflingen vorgelegt.

Die möglichen Ergebnisse dieses Zufallsexperimentes lauten dann:

- für die Kompetenz: $\Omega_K = \{W, NW\}$ und
- für die Performanz: $\Omega_P = \{A_r, A_f, A_o\}$,

d. h.: für jede Aufgabe gilt: (a) Prüflinge wissen die Antwort oder sie wissen sie nicht und (b) Prüflinge geben eine richtige, eine falsche oder gar keine Antwort. Auf diesen Ergebnisräumen lassen sich nun Wahrscheinlichkeiten für alle interessierenden Ereignisse definieren.

Mit p_W bezeichnen wir die Wahrscheinlichkeit dafür, dass ein Prüfling die Antwort auf die gestellte Aufgabe *weiß*. Vermutlich wird diese Wahrscheinlichkeit für unterschiedliche Aufgaben unterschiedliche Werte annehmen. Wir betrachten sie hier allerdings für jeden einzelnen Prüfling als eine *Konstante*, die das Wissen dieses Prüflings ausdrückt. Ein Wert von $p_W = .70$ könnte dann z. B. interpretiert werden als: der Prüfling weiß (!) die Antwort auf 70% der Aufgaben, beherrscht also 70% des Stoffes. Somit ist p_W also genau die Größe, die der

Prüfende bestimmen möchte. Allerdings kann diese Wahrscheinlichkeit nicht direkt geschätzt werden, weil das Wissen nicht beobachtbar ist.

Empirischen Zugang haben wir nur zu den beobachtbaren Ergebnissen A_r, A_f und A_o. Deren Wahrscheinlichkeiten können über die relativen Häufigkeiten geschätzt werden. Den Zusammenhang mit dem Wissen stellen wir durch die folgenden Annahmen her:

1. Prüflinge, die die Antwort auf eine Aufgabe *wissen*, geben auf jeden Fall eine Antwort. Wenn die Antwort *falsch* ist, sprechen wir von einem „Flüchtigkeitsfehler“ (*careless mistake*). Die Wahrscheinlichkeit dafür nennen wir f und definieren:

   $$f := P(A_f|W).$$

   Die Wahrscheinlichkeit für eine *richtige* Antwort im Zustand des Wissens ist deshalb:

   $$P(A_r|W) = 1 - f.$$

2. Prüflinge, die die Antwort auf eine Aufgabe *nicht wissen*, haben zwei Optionen: sie können raten (die Wahrscheinlichkeit, dass sie dies tun, bezeichnen wir mit h) oder sie können die Antwort verweigern. Letzteres führt zum Ergebnis A_o. Ersteres, das Raten, führt mit der Wahrscheinlichkeit

   $$g := P(A_r|\text{„wählt zufällig eine Antwort“})$$

   zu einer *richtigen* Antwort und mit Wahrscheinlichkeit $1 - g$ zu einer *falschen* Antwort. Die mit g bezeichnete Wahrscheinlichkeit, beim Raten die richtige Antwort auszuwählen ist die sogenannte Ratewahrscheinlichkeit. Die Wahrscheinlichkeit zu raten ist die hier mit h bezeichnete Rateneigung.

Aus diesen Annahmen, die in Abbildung 2.1 als Entscheidungsbaum grafisch veranschaulicht sind, lassen sich nun die Wahrscheinlichkeiten für die drei beobachtbaren Ereignisse in Abhängigkeit vom Wissen des Prüflings p_W und den drei Parametern g, f und h ausdrücken. Es gilt nämlich:

$$\begin{aligned} P(A_r) &= p_W \cdot (1 - f) + (1 - p_W) \cdot h \cdot g \\ &= p_W \cdot (1 - f - h \cdot g) + h \cdot g \end{aligned} \tag{2.1}$$

Gleichung 2.1 enthält das wichtigste Ergebnis dieses Kapitels. Sie zeigt nämlich, dass unter den

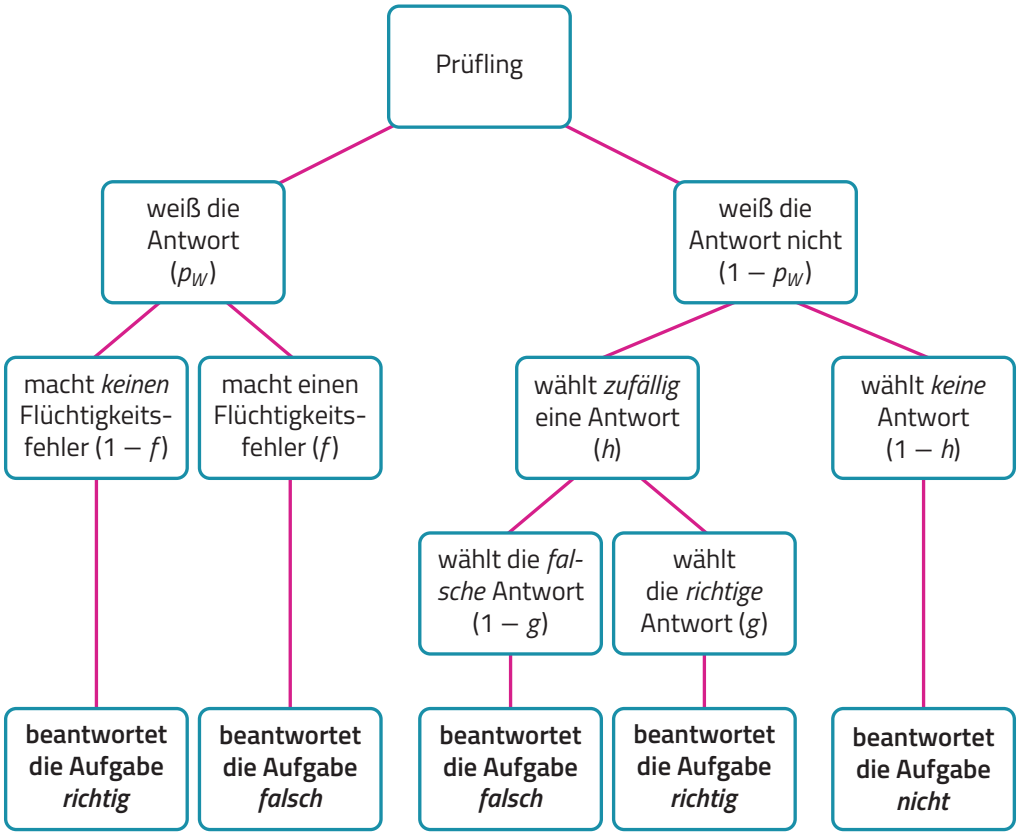


Abbildung 2.1. Ein einfaches Modell für die Bearbeitung von Aufgaben nach dem Antwort-Wahl-Verfahren. In Klammern sind jeweils die Wahrscheinlichkeiten für die entsprechenden Ereignisse angegeben.

oben getroffenen Annahmen die Wahrscheinlichkeit für eine richtige Antwort linear abhängt vom Wissen des Prüflings und dass die genaue Form dieses linearen Zusammenhangs durch die Parameter g, f und h bestimmt wird. In Abbildung 2.2 wird dieser Zusammenhang an zwei Beispielen erläutert.

Die drei Parameter g, f und h haben jeweils eine eindeutige inhaltliche Interpretation. Bei g handelt es sich um die sogenannte Ratewahrscheinlichkeit, die bei Antwort-Wahl-Klausuren eine so wichtige Rolle spielt. Sie hängt in den meisten Fällen von der Anzahl der vorgegebenen Antwortalternativen bzw. Antwortmöglichkeiten m ab. In einfachen Fällen gilt die reziproke

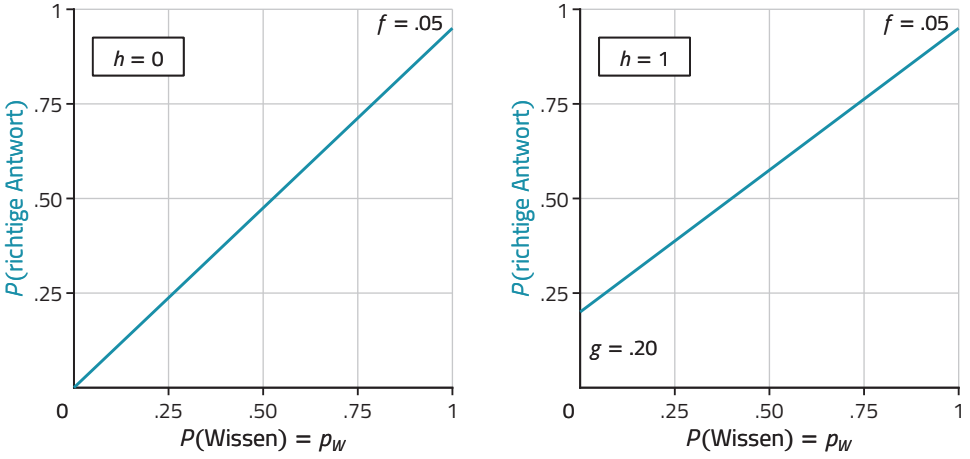


Abbildung 2.2. Die Wahrscheinlichkeit für eine richtige Antwort hängt linear ab von der Wahrscheinlichkeit p_W dafür, dass ein Prüfling die Antwort weiß: $P(\text{richtige Antwort}) = p_W \cdot (1 - f - g \cdot h) + g \cdot h$. Dargestellt ist dies beispielhaft für eine Aufgabe mit einer Ratewahrscheinlichkeit von $g = .20$ und einer Wahrscheinlichkeit für eine falsche Antwort trotz Wissens von $f = .05$. Die Abbildung auf der linken Seite gibt den Verlauf für eine Rateneigung von $h = 0$ an. Auf der rechten Seite beträgt die Rateneigung $h = 1$, so dass selbst ohne jedes Wissen eine Chance in Höhe der Ratewahrscheinlichkeit von $g = .20$ besteht, die Aufgabe richtig zu beantworten.

Beziehung:

$$g = \frac{1}{m}.$$

Mit „Raten“ ist also gemeint: aus den m-vielen Antwortalternativen wird *zufällig* eine gewählt (durch Würfeln, Münzwurf oder einen ähnlichen gleichverteilten Auswahlprozess). Prüflinge, die die eine oder andere Antwortalternative ausschließen können (oder manche Antwortalternativen aus inhaltlichen oder auch formalen Überlegungen favorisieren), haben nicht etwa eine höhere Ratewahrscheinlichkeit, sondern verfügen über „partielles Wissen“, das durch p_W ausgedrückt wird. Umgekehrt heißt das: die Ratewahrscheinlichkeit g lässt sich *nicht* (wie gelegentlich behauptet wird) durch sorgfältige Formulierung der Antwortalternativen reduzieren. Sie entspricht immer der Wahrscheinlichkeit dafür, eine richtige Antwort zu produzieren, wenn eine Antwortalternative durch einen gleichverteilten Zufallsprozess ausgewählt wird.

Anders als die Ratewahrscheinlichkeit wird die Wahrscheinlichkeit für einen „Flüchtigkeitsfehler“ f von Prüfenden nur in den seltensten Fällen thematisiert. Dabei ist es aus Gründen

der Fairness unerlässlich, auch diesen „Fehler 2. Art“ zumindest in Erwägung zu ziehen. Die Gründe für eine Falschantwort trotz sicheren Wissens können vielfältig sein: eine momentane Unaufmerksamkeit, eine missverständliche Formulierung, ein versehentliches Verrutschen in der Antwortzeile. Im Gegensatz zur Ratewahrscheinlichkeit ist hier eine numerische Schätzung allerdings kaum möglich. Üblicherweise wird bei Klausuren unausgesprochen ein Wert von $f = 0$ angenommen („kommt nicht vor“). Es wäre unseres Erachtens durchaus angemessen, die Wahrscheinlichkeit für Flüchtigkeitsfehler explizit zu berücksichtigen durch einen Wert von z. B. $f = .05$. Wer den Fehler 1. Art in der Klausurenbewertung berücksichtigt, sollte dies auch für den Fehler 2. Art tun. Das würde im Übrigen auch die Akzeptanz aller Maßnahmen zur Berücksichtigung der Ratewahrscheinlichkeit durch die Prüflinge – und im Konfliktfall durch Gerichte – sicherlich erhöhen..

Die Neigung, im Zweifel zu raten, die durch den Parameter *h* modelliert wird, ist ebenfalls ein wichtiger Parameter, der die Wahrscheinlichkeit für eine richtige Antwort bei Unwissenheit beeinflusst. In der Abbildung 2.2 ist das deutlich erkennbar. Der Parameter *h* wird zum einen von der Persönlichkeit des Prüflings abhängen, z. B. von dessen Risikobereitschaft, zum anderen wird er aber auch stark beeinflusst sein von dem Scoring-Verfahren, das der Prüfer wählt: je stärker eine Falschantwort sanktioniert wird (z. B. durch Maluspunkte), desto geringer dürfte die Neigung sein, im Falle der Unwissenheit „blind“ zu raten. Die motivationalen Aspekte beim Raten und ihre Abhängigkeit vom Scoring-Verfahren werden daher in den nächsten Kapiteln noch eine wichtige Rolle spielen.

3 Scoring

Punkte für richtige und falsche Antworten

Für richtige Antworten werden üblicherweise Punkte vergeben. Die Gesamtzahl der von den Prüflingen erreichten Punkte bestimmt dann die Note. Bei der Vergabe von Punkten für unterschiedliche Antworten, dem sogenannten Scoring, wird häufig versucht, die Ratewahrscheinlichkeit zu kompensieren, z. B. durch negative Punktwerte für falsche Antworten oder dadurch, dass Punkte nur dann vergeben werden, wenn *alle* Aufgaben einer Aufgabengruppe richtig beantwortet wurden.

Im Wesentlichen gibt es drei Gruppen von Scoringverfahren, die in den nachfolgenden Abschnitten genauer betrachtet werden sollen:

Punkte für richtige Antworten. Für jede richtige Antwort gibt es einen (oder auch mehrere) Punkt(e). Für falsche oder fehlende Antworten gibt es keine Punkte, es werden aber auch keine Punkte abgezogen. Dieses Verfahren wird im Folgenden als Standardverfahren bezeichnet und in Abschnitt 3.1 genauer beschrieben.

Maluspunkte. Im Unterschied zum Standardverfahren werden hier für falsche Antworten Punkte abgezogen. Das ist bei vielen Prüfenden beliebt, wird gelegentlich (z. B. bei nur zwei Antwortalternativen) sogar als unverzichtbar bezeichnet: „Teilbewertung und Punktabzug bei Auswahl der falschen Alternative sind hier obligatorisch“ (Zentrum für Multimedia in der Lehre, o. D., beim Stichwort „Antwortpaare“). Es ist jedoch juristisch problematisch (z. B. OVG Nordrhein-Westfalen, 2008; VG Arnsberg, 2012; s. a. Ludwig, 2014) und wird von Studierenden in der Regel als unfair wahrgenommen. Wir klären in Abschnitt 3.2 die testtheoretischen Wurzeln, diskutieren Vor- und Nachteile und zeigen, dass Maluspunkte immer vermeidbar sind und wie man sie korrekt anwendet, wenn man das trotz ihrer Problematik tun will.

Testlet Scoring. Bei diesen Verfahren werden Aufgaben zu einer Aufgabengruppe zusammengefasst und Punkte in der Regel erst dann vergeben, wenn *alle* Aufgaben einer Aufgabengruppe richtig beantwortet wurden. Auch hier gibt es Variationen, z. B. halbe Punkte, wenn alle Aufgaben bis auf eine richtig beantwortet wurden. Die Konsequenzen dieser Scoringverfahren werden in Abschnitt 3.3 diskutiert.

3.1 Punkte für richtige Antworten: Das Standardverfahren

Beim Scoring wird den Ergebnissen „richtige Antwort", „falsche Antwort" und „keine Antwort" des Zufallsexperiments aus Abschnitt 2.2 jeweils ein Punktwert zugeordnet. Damit ist für jede Aufgabe i eine Zufallsvariable X_i definiert, die im einfachsten Fall des Standardverfahrens eine binäre Zufallsvariable mit einer Bernoulli-Verteilung darstellt:

$$X_i := \begin{cases} 1 & \text{falls Antwort auf Aufgabe } i \text{ richtig ist } (A_r) \\ 0 & \text{sonst } (A_f \text{ oder } A_0). \end{cases}$$

D. h., X_i ist eine Bernoulli-Variable mit $p := p_W \cdot (1 - f - h \cdot g) + h \cdot g$ für alle i. Erwartungswert und Varianz der Zufallsvariablen X_i sind gegeben durch:

$$\mathcal{E}(X_i) = p \quad \text{und}$$
$$\text{var}(X_i) = p \cdot (1 - p).$$

Für einen Test mit n-vielen Aufgaben gleicher Wahrscheinlichkeit p ist die Gesamtanzahl der im Test erreichten Punkte wieder eine Zufallsvariable:

$$X := \sum_{i=1}^{n} X_i.$$

X ist binomialverteilt mit:

$$P(X = r) = \binom{n}{r} \cdot p^r \cdot (1 - p)^{n-r} \tag{3.1}$$

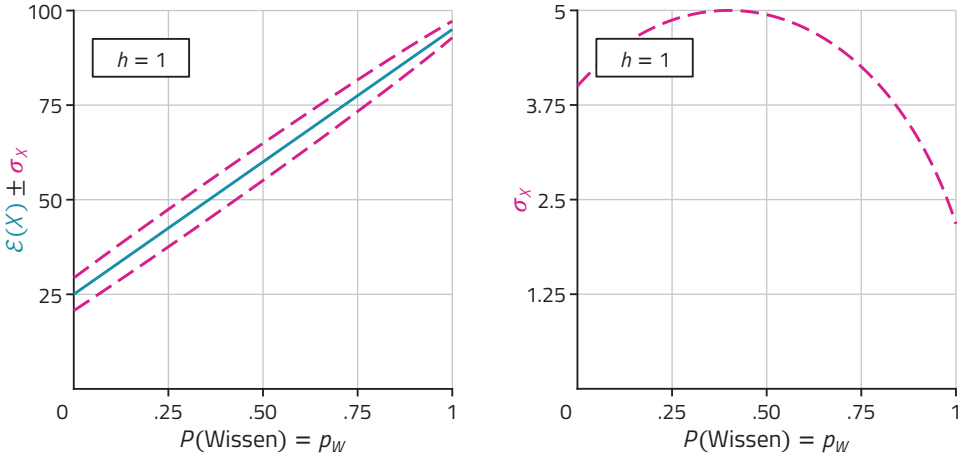


Abbildung 3.1. Links: Verlauf des Erwartungswertes (blaue Linie) für das Klausurergebnis in Abhängigkeit vom Wissen p_W beim Standardscoring für eine Klausur mit $n = 100$ Aufgaben ($g = .25$, $h = 1$, $f = .05$). Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an.
Rechts: Abhängigkeit der Standardabweichung von X von p_W für das nebenstehende Beispiel.

und es gilt:

$$\mathcal{E}(X) = n \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot p \cdot (1 - p).$$

Das bedeutet: Die Verteilung der in einer Klausur erreichten Punktwerte hängt nur ab von der Wahrscheinlichkeit für eine richtige Antwort p (und der Anzahl der Aufgaben im Test). Die Wahrscheinlichkeit p wiederum hängt ab vom Wissen des Prüflings (p_W), von dessen Neigung zu raten h, von der Ratewahrscheinlichkeit g und der Wahrscheinlichkeit für einen Flüchtigkeitsfehler f.

Damit sind die Auswirkungen dieser Parameter auf das Klausurergebnis exakt darstellbar. Wir verwenden dafür hier und in den folgenden Kapiteln Grafiken, bei denen der Erwartungswert des Klausurergebnisses X in Abhängigkeit vom Wissen des Prüflings p_W dargestellt wird. Zur einfacheren Interpretation sind die Beispiele in der Regel so gewählt, dass maximal 100 Punkte im Test erreicht werden können. Abbildung 3.1 zeigt den Erwartungswert für das Testergebnis X in einem Test mit 100 *single-response*-Aufgaben mit jeweils vier Antwortalternativen und genau einer richtigen Antwort. Die Ratewahrscheinlichkeit beträgt $g = 1/4$, die

Rateneigung ist mit $h = 1$ angenommen und die Wahrscheinlichkeit für einen Flüchtigkeitsfehler mit $f = .05$. Der Erwartungswert für X hängt linear ab vom Wissen p_W. Prüflinge, die *nichts* wissen ($p_W = 0$) und deshalb bei jeder Aufgabe blind raten, erreichen im Durchschnitt 25 der möglichen 100 Punkte. Prüflinge, die *alles* wissen ($p_W = 1$), erreichen im Durchschnitt dennoch nur 95 Punkte, da sie wegen der angenommenen Flüchtigkeitsfehlerwahrscheinlichkeit von 5% fünf Punkte bei Aufgaben vergeben, die sie tatsächlich gewusst haben. Zwischen diesen beiden Endpunkten steigt der Erwartungswert für das Testergebnis linear mit dem Wissen an. Natürlich gibt es eine Zufallsschwankung, die durch die Standardabweichung σ_X quantifiziert wird und in der Abbildung durch pinkfarbene, gestrichelte Linien gekennzeichnet ist.

Für den einfachen Fall des Standard-Scorings enthält das Diagramm in Abbildung 3.1 keine großen Überraschungen oder Erkenntnisse. Es ist aber hervorragend geeignet, die Auswirkungen von Wissen, Ratewahrscheinlichkeit, Scoringverfahren usw. auf die zu erwartenden Testergebnisse zu untersuchen. Wir werden dieses Tool deshalb im Folgenden verwenden, um zu zeigen, welche Auswirkungen die unterschiedlichen Ratewahrscheinlichkeiten verschiedener Aufgabenformate, verschiedene Scoringverfahren, die Länge eines Tests und andere Parameter auf die Verteilung der Testergebnisse haben. Auf dieser Grundlage lassen sich dann angemessene Begründungen für die optimale Gestaltung von AW-Klausuren und ihre Auswertung ableiten.

3.2 Maluspunkte

3.2.1 Naive Vergabe von Maluspunkten

Die einfachste und in vielen Fällen naiv angewandte Praxis von Maluspunkten lautet schlicht: Einen Punkt für jede richtige Antwort, einen Minuspunkt für jede falsche Antwort und kein Punkt für fehlende Antworten. Die Zufallsvariable X_i ist damit definiert durch:

$$X_i := \begin{cases} 1 & \text{falls Antwort auf Aufgabe } i \text{ richtig ist } (A_r) \\ -1 & \text{falls Antwort auf Aufgabe } i \text{ falsch ist } (A_f) \\ 0 & \text{sonst (keine Antwort } A_0). \end{cases}$$

Die Wahrscheinlichkeiten für die drei möglichen Ergebnisse sind nach dem Modell aus Abschnitt 2.2 definiert durch:

$$
\begin{aligned}
P(A_r) &= p_W \cdot (1 - f - h \cdot g) + h \cdot g, \\
P(A_f) &= p_W \cdot f + (1 - p_W) \cdot h \cdot (1 - g) \\
&= p_W \cdot (f + h \cdot g - h) + h \cdot (1 - g) \quad \text{und} \\
P(A_0) &= (1 - p_W) \cdot (1 - h) \\
&= p_W \cdot (h - 1) + 1 - h.
\end{aligned}
\tag{3.2}
$$

Daraus lässt sich der Erwartungswert der Zufallsvariablen X_i für das naive Maluspunkte-Scoring berechnen durch:

$$
\begin{aligned}
\mathcal{E}(X_i) &= P(A_r) - P(A_f) \\
&= p_W \cdot (1 - 2f + h - 2 \cdot h \cdot g) - h + 2 \cdot h \cdot g.
\end{aligned}
$$

Für eine Klausur mit n-vielen Aufgaben gleicher Wahrscheinlichkeiten $P(A_r)$ und $P(A_f)$ ist die Gesamtanzahl der im Test erreichten Punkte wieder die Zufallsvariable:

$$
X := \sum_{i=1}^{n} X_i \quad \text{mit} \quad \mathcal{E}(X) = n \cdot \mathcal{E}(X_i).
$$

Die grafische Darstellung der Abhängigkeit des Erwartungswertes für die Gesamtpunktzahl in der Klausur vom Wissen p_W in Abbildung 3.2 zeigt die Konsequenzen des naiven Maluspunkt-Scorings:

- Maluspunkte sind außerordentlich wirksam, wenn verhindert werden soll, dass Punkte nur durch Raten erreicht werden. Prüflinge, die nichts wissen ($p_W = 0$), können im Durchschnitt nicht mit Punkten rechnen. Im Gegenteil: der Erwartungswert liegt im Allgemeinen sogar im negativen Bereich. Das ist auch beabsichtigt – zumindest scheint es auf den ersten Blick so zu sein.
- Maluspunkte führen auch dazu, dass blindes Raten keine rationale Option ist, wenn man die Antwort nicht weiß: Je höher die Rateneigung h ist, desto geringer ist der Erwartungswert. Auch das ist von vielen Prüfenden vielleicht so beabsichtigt.

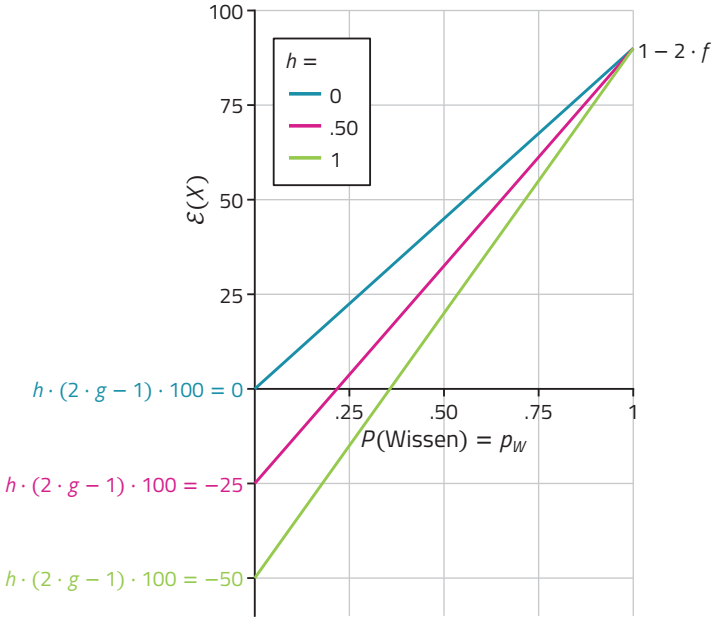


Abbildung 3.2. Verlauf des Erwartungswertes für das Klausurergebnis in Abhängigkeit vom Wissen p_W beim naiven Maluspunkte-Scoring für eine Klausur mit $n = 100$ Aufgaben ($g = .25$, $f = .05$). Die drei Geraden entsprechen drei verschiedenen Werten für die Rateneigung: $h = 0$, $h = .50$ und $h = 1$.

Aber:

- Die Ratekorrektur schießt – zumindest in unserem Beispiel in Abbildung 3.2 – weit über das Ziel hinaus. Prüflinge, die wenig wissen und gelegentlich raten, erhalten eine negative Punktzahl und es ist völlig unklar, was das bedeutet.
- Die starke Sanktionierung von falschen Antworten führt dazu, dass Raten entmutigt wird und Antworten nur dann gegeben werden, wenn die Prüflinge ihrer Sache sicher sind. Das ist aber stark von Persönlichkeitsvariablen abhängig (z. B. Budescu & Bo, 2015), d. h.: Personen die ängstlich sind oder risikoscheu, werden vor allem dann benachteiligt, wenn sie viel wissen.
- Dazu kommt, dass Flüchtigkeitsfehler zusätzlich verschärft werden: Prüflinge, die alles wissen ($p_W = 1$), aber bei nur 5% der Aufgaben einen Fehler machen, haben am Ende nur 90% der erreichbaren Punkte.

Insgesamt gesehen hat das naive Maluspunktsystem gravierende (und den meisten Prüfenden nicht bewusste) Nachteile: Die bei gegebenem Wissen erreichte Punktzahl hängt stark von der persönlichen Rateneigung ab. Sie hängt auch von der Ratewahrscheinlichkeit ab, jetzt aber so, dass das zu erwartende Testergebnis automatisch umso geringer ist, je kleiner die Ratewahrscheinlichkeit ist. Und schließlich wird die Auswirkung des Fehlers zweiter Art, der Flüchtigkeitsfehler, zusätzlich erhöht.

3.2.2 *Formula Scoring*

In der Testtheorie werden Maluspunkte vor allem im Zusammenhang mit dem sogenannten *formula scoring* diskutiert (Holzinger, 1924; Lord & Novick, 1968; Lord, Novick & Birnbaum, 2008). Auch hier werden für Falschantworten Punkte abgezogen – allerdings ist der Punktabzug abhängig von der Ratewahrscheinlichkeit. Bei Aufgaben mit hohen Ratewahrscheinlichkeiten werden mehr Punkte abgezogen als bei Aufgaben mit geringer Ratewahrscheinlichkeit.

Exakt formuliert heißt das: Die Zufallsvariable X_i ist beim *formula scoring* definiert durch:

$$X_i := \begin{cases} 1 & \text{falls Antwort auf Aufgabe } i \text{ richtig ist } (A_r) \\ -g/(1-g) & \text{falls Antwort auf Aufgabe } i \text{ falsch ist } (A_f) \\ 0 & \text{sonst (keine Antwort } A_0). \end{cases}$$

Die Wahrscheinlichkeiten für die drei möglichen Ergebnisse A_r, A_f und A_0 sind nach dem Modell aus Abschnitt 2.2 unverändert definiert durch Gleichung 3.2, so dass der Erwartungswert der Zufallsvariablen X_i für das *formula scoring* berechnet wird mit:

$$\mathcal{E}(X_i) = P(A_r) - \frac{g}{1-g} \cdot P(A_f).$$

Nach Einsetzen der Werte für $P(A_r)$ und $P(A_f)$ aus Gleichung 3.2 und einigen elementaren algebraischen Umformungen erhält man daraus als Ergebnis:

$$\mathcal{E}(X_i) = p_W \cdot \left(1 - \frac{f}{1-g}\right).$$

Für eine Klausur mit n-vielen Aufgaben gleicher Wahrscheinlichkeiten $P(A_r)$ und $P(A_f)$ ist die

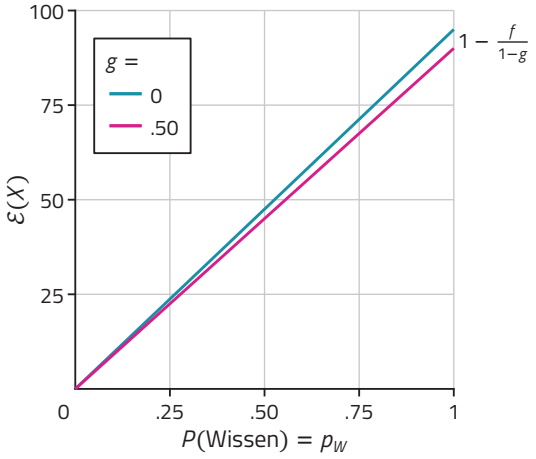


Abbildung 3.3. Verlauf des Erwartungswertes für das Klausurergebnis in Abhängigkeit vom Wissen p_W beim *formula scoring* für eine Klausur mit $n = 100$ Aufgaben ($f = .05$). Die Geraden sind unabhängig von h und entsprechen zwei verschiedenen Werten für die Ratewahrscheinlichkeit: $g = 0$ und $g = .50$.

Gesamtanzahl der in der Klausur erreichten Punkte wieder die Zufallsvariable:

$$X := \sum_{i=1}^{n} X_i \quad \text{mit} \quad \mathcal{E}(X) = n \cdot \mathcal{E}(X_i).$$

Die grafische Darstellung der Abhängigkeit des Erwartungswertes für die Gesamtpunktzahl in der Klausur vom Wissen p_W in Abbildung 3.3 zeigt die Idee beim *formula scoring*: Die Gewichte für die Maluspunkte werden so gewählt, dass sie die Ratewahrscheinlichkeit optimal korrigieren. Insbesondere gilt:

- Der Erwartungswert für die Punkte in der Klausur ist *nicht* mehr abhängig von der Rateneigung h. Prüflinge, die eine Antwort nicht wissen, haben immer einen Erwartungswert von null, sowohl beim Raten als auch beim Auslassen der Antwort. Risikoneigung, Ängstlichkeit und ähnliche Persönlichkeitsvariablen beeinflussen nicht mehr das Testergebnis.
- Auch die Höhe der Ratewahrscheinlichkeit spielt (fast) keine Rolle mehr. Der Fehler erster Art wird durch die Maluspunkte passgenau korrigiert. Allerdings wird der Fehler zweiter Art nicht nur vernachlässigt, sondern sogar verstärkt um den Faktor $1/(1-g)$.
- Die Ratekorrektur beim *formula scoring* führt dazu, dass der Erwartungswert der erreichten Punkte exakt dem Wissen des Prüflings entspricht, wenn Flüchtigkeitsfehler nicht

auftreten ($f = 0$). Ist $f > 0$, dann wird das Wissen tendenziell unterschätzt, und zwar umso mehr, je größer das Wissen ist und je größer die Ratewahrscheinlichkeit ist.

- Juristische Bedenken gegen den Abzug von Punkten, wie sie im Urteil des OVG Nordrhein-Westfalen (2008) zum Ausdruck kommen, betreffen vermutlich auch das *formula scoring*.
- Unschön ist auch, dass zwar die Erwartungswerte alle positiv sind, im Einzelfall für einen Prüfling aber durchaus auch negative Werte als Gesamtergebnis vorkommen können.

Als Fazit ist festzuhalten: Wer Maluspunkte als „Heilmittel" gegen Ratewahrscheinlichkeiten einsetzen will, der sollte das nach den Regeln des *formula scoring* tun und nicht einfach naiv für Falschantworten einen Punkt abziehen. Die juristischen Probleme beim Punktabzug sind damit vermutlich nicht gelöst. Und auch in der Testtheorie sind die Konsequenzen des *formula scoring* für Reliabilität und Validität umstritten (Lord, 1975).

3.3 *Testlet Scoring*

Eine ganz andere Strategie beim Umgang mit hohen Ratewahrscheinlichkeiten bei AW-Klausuren ist es, die Schwelle für die Vergabe von Punkten höher zu setzen. Dabei sind für die Beantwortung einer Aufgabengruppe mehrere Einzelantworten erforderlich. Punkte werden erst dann gutgeschrieben, wenn alle (oder in manchen Variationen: fast alle) Einzelaufgaben richtig beantwortet wurden. In der Testtheorie wird häufig der Begriff *testlet scoring* verwendet, wenn Punkte nicht separat und unabhängig für jede Einzelaufgabe vergeben werden, sondern nur für eine Gruppe von (Teil-)Aufgaben insgesamt (Wainer & Kiely, 1987). Wir beschreiben in diesem Kapitel die wichtigsten *testlet-scoring*-Verfahren für AW-Klausuren, definieren wieder X als Zufallsvariable, die die Gesamtanzahl der Punkte in einer Klausur repräsentiert und betrachten wie in den vorhergehenden Kapiteln $\mathcal{E}(X)$ als Funktion des Wissens p_W.

3.3.1 Alles-oder-nichts

Bei diesem Scoringverfahren gibt es einen Punkt nur für eine vollständige und fehlerfreie Lösung aller Teilaufgaben. Bei einem oder mehreren Fehlern werden 0 Punkte vergeben. Um diese Situation zu formalisieren, denken wir uns jede Aufgabe i zusammengesetzt aus k-vielen Teilaufgaben, von denen jede richtig oder falsch beantwortet werden kann.

Bezeichnen wir mit Y_{ij} die binäre Indikatorvariable für das Lösen der Teilaufgaben, heißt das:

$$Y_{ij} := \begin{cases} 1 & \text{falls Antwort auf Aufgabe } i \text{ Teil } j \text{ richtig ist } (A_r) \\ 0 & \text{falls Antwort auf Aufgabe } i \text{ Teil } j \text{ falsch ist } (A_f) \text{ oder fehlt } (A_0). \end{cases}$$

Die Summe der richtigen Teilantworten ergibt sich als:

$$Y_i := \sum_{j=1}^{k} Y_{ij}$$

und für das Alles-oder-nichts-Scoring gilt:

$$X_i := \begin{cases} 1 & \text{falls } Y_i = k \\ 0 & \text{sonst.} \end{cases}$$

Die Wahrscheinlichkeiten $P(A_r)$ und $P(A_f)$ aus Gleichung 3.2 beziehen wir auf jede einzelne Teilaufgabe. X_i ist damit für jede Aufgabe i eine Bernoulli-Variable mit:

$$\begin{aligned} \mathcal{E}(X_i) &= P(A_r)^k \quad \text{und} \\ \text{var}(X_i) &= P(A_r)^k \cdot (1 - P(A_r)^k). \end{aligned}$$

Das Gesamtergebnis einer Klausur mit n-vielen Aufgaben fassen wir wieder als Zufallsvariable auf:

$$X := \sum_{i=1}^{n} X_i$$

und es gilt: X ist binomialverteilt mit:

$$\begin{aligned} \mathcal{E}(X) &= n \cdot P(A_r)^k \quad \text{und} \\ \text{var}(X) &= n \cdot P(A_r)^k \cdot (1 - P(A_r)^k). \end{aligned}$$

Was das bedeutet, veranschaulicht Abbildung 3.4.

- Die Wahrscheinlichkeit, Punkte durch zufälliges Raten zu erlangen, ist tatsächlich extrem reduziert. Prüflinge, die wenig wissen, haben einen Erwartungswert nahe null und mit zunehmendem Wissen steigt der Erwartungswert nur langsam an.

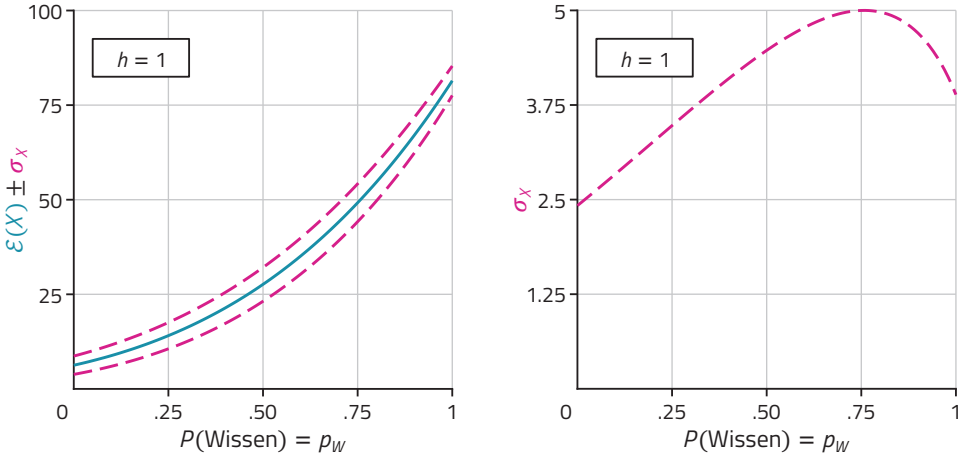


Abbildung 3.4. Links: Verlauf des Erwartungswertes (blaue Linie) für das Klausurergebnis in Abhängigkeit vom Wissen p_W beim Alles-oder-nichts-Scoring für eine Klausur mit $n = 100$ Aufgaben, die jeweils aus $k = 4$ Teilaufgaben bestehen. Für jede Teilaufgabe sei $g = 1/2$ und $f = .05$. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an.
Rechts: Abhängigkeit der Standardabweichung von X von p_W für das nebenstehende Beispiel.

- Allerdings heißt das auch: Prüflinge, die einen Erwartungswert von 50% der maximalen Punktzahl erreichen wollen, müssen über 75% des Stoffes beherrschen ($p_W = .7576$).
- Selbst Prüflinge, die alles wissen ($p_W = 1$), haben einen Erwartungswert von lediglich 81.5% der maximalen Punktzahl. Das liegt im Beispiel natürlich an der Flüchtigkeitsfehlertoleranz ($f = .05$), die sich beim Alles-oder-nichts-Scoring dramatisch auswirkt.
- Der Zusammenhang zwischen erwartetem Ergebnis und Wissen ist nicht mehr linear.
- Die Varianz und damit die Standardabweichung von X ist besonders groß bei einem Wissen im Bereich von $p_W = .75$.

Die angestrebte Reduktion der Ratewahrscheinlichkeit ist also teuer erkauft. Sie geht vor allem auf Kosten eines enorm erhöhten Fehlers zweiter Art. Darüber hinaus sind die hohen Streuungen im Kontext von Klausuren ein erhebliches Problem. Hohe Streuung bedeutet: Zwei Prüflinge mit gleich umfangreichem Wissen p_W müssen damit rechnen, dass sich ihre Klausurergebnisse stark unterscheiden.

Betrachtet man die Standardabweichung von X in Abhängigkeit von p_W (Abbildung 3.4, rechts), stellt man fest, dass sie vor allem in dem Bereich am größten ist, in dem vermutlich die meisten Prüflinge liegen werden. Das verschärft das Problem zusätzlich. Der Grund für diesen

Verlauf ist leicht zu verstehen: Prüflinge, die wenig wissen, haben kaum Chancen, Punkte zu bekommen; es gibt wenig Spielraum für Variabilität. Bei Prüflingen, die viel wissen, aber doch nicht alles, ist die Variabilität am größten. Es kommt darauf an, ob die Wissenslücken quer über alle Aufgaben verteilt sind, oder ob sie sich auf einzelne Aufgaben konzentrieren. Im ersten Fall können nur wenige Punkte erreicht werden, im letztgenannten relativ viele.

3.3.2 K' bzw. „Kprim"

Wegen der offensichtlichen Probleme des Alles-oder-nichts-Scorings, die in der praktischen Anwendung durch ungewöhnlich schlechte Klausurergebnisse auffallen, wird gelegentlich ein etwas abgeschwächtes Verfahren empfohlen, das vor allem in der Medizin unter der Bezeichnung K' (oder auch „Kprim") bekannt ist (z. B. Krebs, 2004). In der Standardvariante besteht hier jede Aufgabe aus vier binären Teilaufgaben. Sind alle vier Teilaufgaben richtig beantwortet, wird ein Punkt für die Aufgabe vergeben, bei drei richtigen Teilaufgaben ein halber Punkt. Bei weniger als drei richtigen Teilaufgaben werden 0 Punkte vergeben.

Für unsere Zufallsvariable bedeutet das (die Anzahl der Teilaufgaben k ist hier mit vier festgelegt, für andere Werte lässt sich die Rechnung natürlich analog durchführen):

$$X_i := \begin{cases} 1 & \text{falls } Y_i = 4 \\ 1/2 & \text{falls } Y_i = 3 \\ 0 & \text{sonst.} \end{cases}$$

Daraus ergibt sich als Erwartungswert für jede einzelne Aufgabe i:

$$\begin{aligned} \mathcal{E}(X_i) &= p^4 + 1/2 \cdot \binom{4}{3} \cdot p^3 \cdot (1-p) \\ &= p^4 + 2p^3 \cdot (1-p) \\ &= 2p^3 - p^4 \end{aligned}$$

mit:

$$p := P(A_r) = p_W \cdot (1 - f - h \cdot g) + h \cdot g.$$

Die Varianz von X_i lässt sich am einfachsten über den Verschiebungssatz

$$\text{var}(X) = \mathcal{E}(X^2) - (\mathcal{E}(X))^2$$

berechnen. Wegen

$$\begin{aligned}\mathcal{E}(X_i^2) &= p^4 + 1/4 \cdot \binom{4}{3} \cdot p^3 \cdot (1-p) \\ &= p^4 + p^3 \cdot (1-p) \\ &= p^3\end{aligned}$$

gilt für die Varianz von X_i für alle i:

$$\begin{aligned}\text{var}(X_i) &= \mathcal{E}(X_i^2) - (\mathcal{E}(X_i))^2 \\ &= p^3 - (2p^3 - p^4)^2 \\ &= p^3 - 4p^6 + 4p^7 - p^8.\end{aligned}$$

Für das Gesamtergebnis der Klausur mit n-vielen Aufgaben gilt wieder:

$$\begin{aligned}\mathcal{E}(X) &= n \cdot \mathcal{E}(X_i) \quad \text{und} \\ \text{var}(X) &= n \cdot \text{var}(X_i).\end{aligned}$$

Vorausgesetzt wird dabei, dass Auswahl und Beantwortung der Aufgaben (und der Teilaufgaben) stochastisch voneinander unabhängig sind.

Das Ergebnis ist in Abbildung 3.5 wiedergegeben. Für die Rateneigung kann hier $h = 1$ angenommen werden, da „im Zweifel Raten" die erkennbar beste Strategie ist. Abbildung 3.5 zeigt ein sehr ähnliches Bild wie Abbildung 3.4, allerdings in abgeschwächter Form. Das heißt, das K'-Scoring hat alle Eigenschaften des Alles-oder-nichts-Scorings, allerdings sind die Nachteile etwas abgemildert. Prüflinge, die die Hälfte wissen ($p_W = .50$) haben einen Erwartungswert von etwa 50% der erreichbaren Punkte – das klingt vernünftig und führt vermutlich dazu, dass es in der Praxis keine auffälligen Klausurergebnisse gibt. Aber: Prüflinge, die nichts wissen, haben einen Erwartungswert von fast 20%. Dagegen kommen Prüflinge, die alles wissen, lediglich auf etwa 90% der Punkte, der Verlauf ist nicht linear und die Standardabweichung ist relativ groß.

Aus juristischer Sicht scheint übrigens sowohl das Alles-oder-nichts-Scoring als auch K'-Scoring unproblematisch zu sein (OVG Nordrhein-Westfalen, 2016, Rn. 4 & 6ff). Und schließlich wird eine der bekanntesten AW-Klausuren, die deutsche Führerscheinprüfung, nach dem Alles-oder-nichts-Verfahren bewertet.

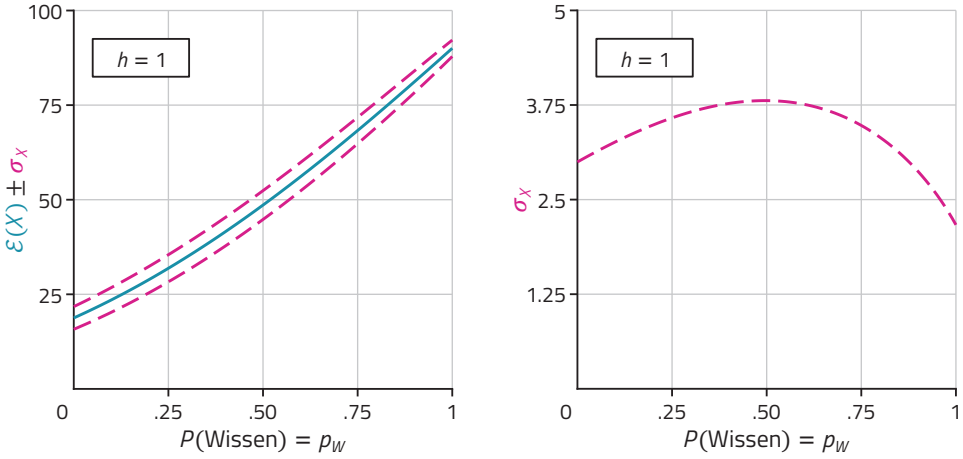


Abbildung 3.5. Links: Verlauf des Erwartungswertes (blaue Linie) für das Klausurergebnis in Abhängigkeit vom Wissen p_W beim K'-Scoring für eine Klausur mit $n = 100$ Aufgaben, die jeweils aus $k = 4$ Teilaufgaben bestehen. Für jede Teilaufgabe sei $g = 1/2$ und $f = .05$. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an.
Rechts: Abhängigkeit der Standardabweichung von X von p_W für das nebenstehende Beispiel.

4 Bestehenskriterium und Notenvergabe

Bei einer Klausur ist die erreichte Punktzahl im Allgemeinen die Grundlage für das Bestehen oder Nichtbestehen der Klausur bzw. für die in der Klausur erreichte Note als Leistungsbewertung. Die Bestehens- und Notengrenzen werden entweder von Prüfungsordnungen vorgegeben oder von den Prüfenden im Einzelfall festgelegt. Ein sehr häufig verwendetes Schema ist in beiden Fällen die in Tabelle 4.1 angegebene Zuordnung von Punktwerten zu Noten. Bestanden haben nach diesem Schema Prüflinge, die mindestens 50% der maximalen Punkte erreicht haben, für ein „sehr gut" (1.0) sind mindestens 95% erforderlich, und dazwischen wird linear angestuft. Natürlich sind je nach Prüfungsanforderung oder Ausbildungsziel auch andere Bestehens- und Notengrenzen denkbar. Jedes starre und fest definierte Benotungsschema wird aber ad absurdum geführt, wenn es sich auf Klausuren mit unterschiedlicher Ratewahrscheinlichkeit bezieht. Eine Bestehensgrenze von 50% der Maximalpunktzahl mag sinnvoll

Tabelle 4.1. Beispiel für eine typische Zuordnung von Noten zu Punktwerten.

% Maximalpunktzahl		Note
unter	50	nicht bestanden
ab	50	4.0
ab	55	3.7
ab	60	3.3
ab	65	3.0
ab	70	2.7
ab	75	2.3
ab	80	2.0
ab	85	1.7
ab	90	1.3
ab	95	1.0

und vertretbar sein bei Klausuren ohne nennenswerte Ratewahrscheinlichkeit (z. B. bei frei zu formulierenden Antworten, Rechenaufgaben etc.). Bei AW-Aufgaben mit einer Ratewahrscheinlichkeit von $g = .25$ ist das höchst fragwürdig, und bei einer Ratewahrscheinlichkeit von $g = 1/2$ entsprechen 50% bereits dem Erwartungswert für reines Raten ohne jedes Wissen.

Das ist kein Geheimnis. Das Problem der Ratewahrscheinlichkeit ist allgemein bekannt, es wird immer wieder einmal beklagt (z. B. Klein, 2016), aber die praktizierten Konsequenzen beschränken sich in der Regel auf die oben beschriebenen Strategien mit Maluspunkten und erschwerter Punktvergabe (fast immer nach einer wenig reflektierten, intuitiven Logik) – oder auf einen gänzlichen Verzicht auf AW-Aufgaben. Bezeichnend dafür ist ein Bericht von Ludwig (2014) über eine Studentin, die erfolgreich gegen eine Klausurbewertung mit Maluspunkten geklagt hat. Dort heißt es im letzten Satz: „*Ihr Professor bietet heute übrigens keine Multiple-Choice-Klausuren mehr an. Ohne Malus-Punkte* [sic!]*, sagt er, ergebe das einfach keinen Sinn.*" Vielfach wird auch empfohlen, auf bestimmte Formen von AW-Aufgabentypen mit hoher Ratewahrscheinlichkeit zu verzichten. Das ist schade, weil AW-Aufgaben in vielen Bereichen außerordentlich sinnvoll, effektiv und ökonomisch eingesetzt werden können und gerade ihre Vielfalt ein großes Potenzial darstellt. Und es ist unnötig, weil das Problem der Ratewahrscheinlichkeit gut lösbar ist.

Der Kern der Lösung des Problems liegt darin, die Bestehens- und Notengrenzen nicht über die Anzahl der erreichten Punkte zu definieren, sondern über das Wissen der Prüflinge. Statt:

„*Bestanden hat, wer mindestens die Hälfte der Aufgaben richtig* ***beantwortet hat***"

schlagen wir vor, zu formulieren:

„*Bestanden hat, wer mindestens die Hälfte der Aufgaben richtig* ***beantworten kann***"

und damit die Bewertung an der Kompetenz zu orientieren, und nicht an der Performanz. Da das Wissen und die Kompetenz nicht direkt beobachtbar sind, müssen wir uns allerdings in jedem Fall auf die in der Klausur erworbenen Punkte beziehen. Wir können aufgrund unserer Analysen aber für jede beliebige Klausur den Zusammenhang zwischen Wissen und Punkten exakt formulieren. Wenn man eine Notenskala im Sinne von Tabelle 4.1 anstrebt, dann liegt nach diesem Vorschlag die Bestehensgrenze nicht mehr starr bei 50% richtiger Antworten, sondern bei demjenigen Punktwert, der dem Erwartungswert für $p_W = .50$ entspricht. Die Grenze für die Note 1.0 erhält man entsprechend, indem man den Erwartungswert für $p_W = .95$ bildet usw. Ein einfaches Beispiel soll das verdeutlichen.

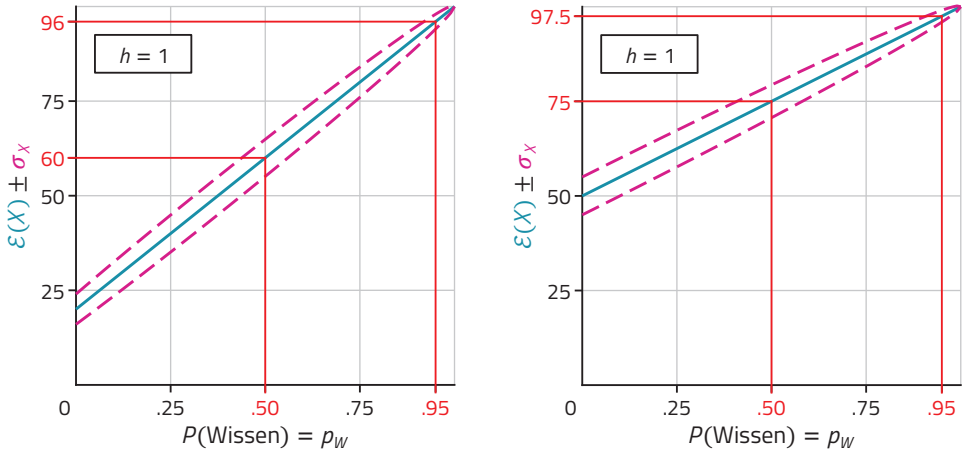


Abbildung 4.1. Zwei Beispiele für Bestehens- und Notengrenzen mit den Ratewahrscheinlichkeiten $g = .20$ (links) und $g = .50$ (rechts). Auf der y-Achse sind die Punktwerte der Bestehensgrenze (Note 4.0), für die Wissen von $p_W = .50$ angesetzt wird, und der Bestnotengrenze (Note 1.0), für die Wissen von $p_W = .95$ angesetzt wird, rot markiert.

Im Staatsexamen Medizin werden *single-response*-Aufgaben mit jeweils fünf Antwortalternativen verwendet, von denen genau eine richtig ist. Unter der Annahme $h = 1$ (Prüflinge, die nichts wissen, raten – das ist die einzig rationale Strategie, und Studierende der Medizin wissen das), $g = .20$ und $f = 0$ (wir sind schließlich in einem Mediziner-Examen) ist der Erwartungswert des Punktwertes für $n = 100$ Aufgaben die in Abbildung 4.1 (links) dargestellte lineare Funktion.

Prüflinge, die die Hälfte des Stoffes beherrschen ($p_W = .50$), werden – im Durchschnitt – 60% der Aufgaben richtig beantworten. Und genau dies ist die Bestehensgrenze im Examen. In dieser Logik ist es auch völlig unproblematisch, binäre Ja/Nein-Aufgaben zu verwenden, und zwar ohne Maluspunkte und mit dem einfachen, unproblematischen Standardscoring aus Abschnitt 3.1. Die Erwartungswertkurve für $h = 1$, $g = .50$ und $f = 0$ in Abbildung 4.1 (rechts) zeigt, dass zum Bestehen einer Klausur, die ausschließlich aus binären Entscheidungen besteht, 75% der Antworten richtig sein müssen und erst bei 97.5% richtiger Antworten die Note 1.0 vergeben wird.

Das Grundprinzip besteht also darin, nicht den individuellen Punktwert eines Prüflings durch Maluspunkte zu korrigieren, sondern die Bestehens- und Notengrenzen an die Ratewahrscheinlichkeiten anzupassen. Das ist transparenter und sowohl Studierenden als auch

Juristen besser zu vermitteln als Punkteabzug. Die Akzeptanz wird spätestens dann endgültig gewonnen, wenn Prüfende konsequent genug sind, nicht nur Ratewahrscheinlichkeiten zu berücksichtigen, sondern auch Flüchtigkeitsfehler. Wer in seinen Klausuren einen Wert von z. B. $f = .05$ oder $f = .01$ als Flüchtigkeitsfehlertoleranz einsetzt, macht sicher nichts falsch und dokumentiert damit Fairness und Auswertungskompetenz.

Für die Praxis ergeben sich daraus die folgenden Empfehlungen:

- Die Klausur wird zusammengestellt nach rein inhaltlichen Gesichtspunkten. Vor- und Nachteile der einzelnen Aufgabentypen und -formate werden in Teil II dieses Handbuches im einzelnen dargelegt. Die Ratewahrscheinlichkeit ist dabei *kein* Güte- oder Auswahlkriterium.
- Als Scoringverfahren wird ausschließlich das Standardscoring verwendet.
- Für die fertige Klausur wird die Funktion F bestimmt, die den Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert für das Klausurergebnis herstellt: $\mathcal{E}(X) = F(p_W)$. Dazu muss man lediglich die Ratewahrscheinlichkeiten g kennen und die Flüchtigkeitsfehlertoleranz f festlegen.
- Der Parameter h kann auf den Wert 1 gesetzt werden, da wir das einfache Standardscoring verwenden und die Prüflinge darauf hinweisen, dass im Zweifel „raten" die beste Strategie ist. Das ist vernünftig, entscheidungstheoretisch korrekt und es sorgt vor allem dafür, dass das Testergebnis nicht abhängt von irrelevanten Persönlichkeitsvariablen wie Risikoneigung, Selbstbewusstsein oder Ängstlichkeit.
- Aus der Funktion F erhalten wir unmittelbar die ratekorrigierten Bestehens- und Notengrenzen für die Klausur. Die Funktion und die Notengrenzen sind unabhängig vom Testergebnis der Klausur und können schon vor dem Klausurtermin bestimmt und gegebenenfalls den Studierenden mitgeteilt werden.

Ist das kompliziert? Keineswegs. Für die Standardformate der AW-Aufgaben sind die Funktionen, wie wir gesehen haben, einfache lineare Funktionen. In Teil II gibt es dazu weitere Beispiele und konkrete Hinweise. Für gemischte Klausuren mit unterschiedlichen Aufgabentypen werden die Funktionen addiert. Das wird im Kapitel 12 ausführlich erläutert. Bei der Verwendung von Prüfungssoftware, wie z. B. EvaExam (Electric Paper Evaluationssysteme GmbH, 2017a) kann die Auswertung automatisch erfolgen, so dass Prüfende sich beim Zusammenstellen der Klausur dazu keine Gedanken machen müssen. Wenn sie wissen, was sie tun, ist das aber keine schlechte Idee. Dazu beizutragen ist das wichtigste Anliegen dieses Handbuches.

5 Aufgabenformate und Aufgabentypen

Im Prüfungsalltag, in der Literatur und in elektronischen Prüfungssystemen finden sich etliche Möglichkeiten zur Gestaltung von Aufgaben. In der an der Martin-Luther-Universität Halle-Wittenberg genutzten E-Learning-Plattform ILIAS stehen z. B. die Aufgabentypen *Single Choice*, *Multiple Choice*, *ImageMap*, Lückentext, Fehlertext, Zuordnungsfrage und einige andere zur Verfügung. Formal betrachtet lassen sich aber fast alle dieser optisch zunächst stark verschieden wirkenden *Aufgabentypen* auf wenige zugrundeliegende *Aufgabenformate* zurückführen. Dabei unterscheiden sich Aufgabenformate voneinander durch die grundsätzlichen Anforderungen, die jedes Format an die Beantwortung einer Aufgabe stellt, während unterschiedliche Aufgabentypen desselben Formats prinzipiell alle gleich funktionieren, sich jedoch in ihrer äußeren Form unterscheiden.

In den in Teil II folgenden Kapiteln werden die wichtigsten Aufgabenformate für die automatisierte Auswertung von elektronischen oder schriftlichen Prüfungen besprochen. Jedem Aufgabenformat ist dabei ein Kapitel gewidmet, in dem jeweils mögliche Scoringverfahren und der Umgang mit der Ratewahrscheinlichkeit beschrieben werden. Für jedes Aufgabenformat wird vorgestellt, wie Wissen und Punkte nach Anwendung des in Kapitel 2 beschriebenen Modells zusammenhängen. Aus diesem Zusammenhang werden auch Empfehlungen für die Ratekorrektur der Bestehens- und Notengrenzen abgeleitet. Außerdem werden Vor- und Nachteile der einzelnen Aufgabenformate zusammengefasst.

Im Mittelpunkt stehen dabei vor allem die verschiedenen Arten von Antwort-Wahl-Aufgaben. Diese gebundenen Aufgaben lassen sich zunächst in drei Arten unterteilen, die sich in Anzahl und formaler Abhängigkeit der Antwortalternativen unterscheiden:

- es ist eine (und nur eine) Antwort zu geben (*single response*, *SR*, s. Kapitel 6),
- es sind mehrere, formal voneinander unabhängige Antworten zu geben (*multiple response*, *MR*) oder

- es sind mehrere, formal voneinander abhängige Antworten zu geben (z. B. Zu- und Anordnungsaufgaben, s. Kapitel 10).

Die *multiple-response*-Aufgaben lassen sich darüber hinaus in zwei Typen unterteilen, die sich in der Anzahl der möglichen Entscheidungsalternativen je Antwortalternative unterscheiden:

- es gibt nur zwei Entscheidungsmöglichkeiten: auswählen oder nicht auswählen (*multiple select*, *MS*, s. Kapitel 7) oder
- es gibt für jede Antwortalternative drei Entscheidungsmöglichkeiten, z. B.: „ja", „nein" und keine Antwort (*multiple true-false*, *MTF*, s. Kapitel 8).

Neben den gebundenen Aufgabenformaten werden in Kapitel 9 offene Aufgabenformate behandelt, da beide Formate in Klausuren häufig gemischt verwendet werden. Kapitel 11 behandelt kurz die sogenannten freien Aufgabenformate, deren Ausgestaltung in der Hand von Prüfenden liegt und sich daher keinem der anderen Formate zuordnen lassen. Auch hier gibt es Hinweise dazu, ob und wie das in Teil I entwickelte probabilistische Modell ggf. angewendet werden kann. Abschließend wird in Kapitel 12 die Ratekorrektur von Bestehens- und Notengrenzen für Klausuren dargestellt, die aus einer Kombination von unterschiedlichen Aufgabenformaten bestehen.

Da dieser Leitfaden insbesondere auf die Gestaltung von Klausuren in ILIAS (ILIAS open source e-Learning e.V., 2017) ausgelegt ist, findet sich in jedem Kapitel auch eine Liste der ILIAS-Aufgabentypen, die dem jeweils besprochenen Aufgabenformat zuzuordnen sind, sowie eine Reihe von Beispielen, die mit ILIAS erstellt wurden. In der nachfolgenden Übersicht in Tabelle 5.1 sind alle ILIAS-Aufgabentypen aufgeführt, welche bei Drucklegung des Handbuchs in ILIAS 5.1 der Martin-Luther-Universität Halle-Wittenberg zur Verfügung standen. Jedem Aufgabentyp sind dabei das Aufgabenformat, dessen Ratewahrscheinlichkeit g, die Seite des entsprechenden Kapitels in diesem Handbuch und die Seite der Kurzreferenz zum Aufgabenformat zugeordnet.

Tabelle 5.1. Alphabetisch sortierte Übersicht über die Aufgabentypen, welche zum Zeitpunkt der Drucklegung dieses Handbuchs in ILIAS 5.1 der Martin-Luther-Universität Halle-Wittenberg zur Verfügung standen. Jedem Aufgabentyp sind das zugehörige Aufgabenformat, die Ratewahrscheinlichkeit g und die Seiten des Kapitels (Kap.) bzw. des *Cheat Sheets* (*CS*) zu diesem Format zugeordnet.

ILIAS-Aufgabentyp	Einordnung	g	Kap. auf Seite	*CS* auf Seite
Anordnungsfrage	abhängige Antwortalternativen	$1/k!$ [a]	89	124
Anordnungsfrage (horizontal)	abhängige Antwortalternativen	$1/k!$ [a]	89	124
Datei hochladen	freies Format	0 [b]	108	126
Fehlertext	*multiple select*	$1/2$ [c]	43	122
Flash-Frage	freies Format	0 [b]	108	126
Formelfrage	offene Aufgabe	0	77	125
Freitext	offene Aufgabe	0	77	125
ImageMap, MC	*multiple select*	$1/2$ [c]	43	122
ImageMap, SC	*single response*	$1/k$	32	121
Java-Applet-Frage	freies Format	0 [b]	108	126
JSME-Frage	offene Aufgabe	0	77	125
Kprim Choice	*multiple true-false*	$1/2$	62	
Long-Menu-Frage	offene Aufgabe	≈ 0 [d]	77	125
Lückentext (Auswahl-Lücke)	*single response*	$1/k$	32	121
Lückentext (numerische Lücke)	offene Aufgabe	0	77	125
Lückentext (Textlücke)	offene Aufgabe	0	77	125
Multiple Choice	*multiple select*	$1/2$ [c]	43	122
numerische Frage	offene Aufgabe	0	77	125
Single Choice	*single response*	$1/k$	32	121
Text-Teilmenge	offene Aufgabe	0	77	125
Zeichenaufgabe	freies Format	0	108	126
Zuordnungsfrage	abhängige Antwortalternativen	– [e]	89	123

[a] bei Anwendung des Alles-oder-nichts-Verfahrens
[b] sofern keines der anderen Aufgabenformate nachgebildet wird
[c] sofern die Antwortalternativen einzeln bewertet werden
[d] sofern die Liste der Begriffe genügend lang ist
[e] dynamisch und abhängig von der Anzahl der Begriffe und dem Scoringverfahren

Teil II

Praktische Anwendung auf gängige Aufgabenformate

6 Das *single-response*-Format

6.1 Charakteristik

Das *single-response-Format* (*SR*-Format) ist sicherlich das bekannteste und am weitesten verbreitete Format für Aufgaben im Antwort-Wahl-Verfahren. In medizinischen Staatsexamina in Deutschland wird bei den geschlossenen Antwortformaten derzeit ausschließlich dieses Format verwendet und auch bei Fragespielen zu Unterhaltungszwecken im Fernsehen oder im Internet scheint sich dieses Format bei den geschlossenen Antwortformaten als Standard durchgesetzt zu haben: Zu einer Frage oder Aufgabenstellung werden *m*-viele Antwortalternativen vorgegeben, von denen genau eine als „richtig" gilt. Alle anderen gelten als „falsch" und werden als Distraktoren bezeichnet. Die Aufgabe des Prüflings besteht darin, die richtige Antwort zu markieren. Daher wird das Format auch als *single-mark*-Format bezeichnet.

Im Alltagssprachgebrauch und auch in einschlägigen Publikationen (z. B. Cronbach, 1939; Kubinger, 2014) wird das *single response*-Format häufig unter dem Oberbegriff *multiple choice* subsumiert . Dies ist jedoch irreführend, da beim *single-response*-Format immer nur genau eine Antwortalternative ausgewählt werden darf und die Auswahl mehrerer Antwortalternativen als falsche Beantwortung der Aufgabe gewertet wird. Im Gegensatz dazu sind beim *multiple-select*-Format, das in Kapitel 7 behandelt wird, auch Mehrfachantworten möglich, da bei diesem Format auch mehrere Antwortalternativen als richtig gelten können. Ob eine Aufgabe mit mehreren Antwortalternativen vom Typ *single response* oder vom Typ *multiple select* ist, kann man ihr oft nicht ohne Weiteres ansehen. Es ist deshalb wichtig, dass das Aufgabenformat den Prüflingen eindeutig kommuniziert wird, da das Format erhebliche Konsequenzen sowohl für die Bearbeitung durch die Prüflinge als auch für die Bewertung durch die Prüfenden hat.

Unter dieser Voraussetzung – die Prüflinge wissen, dass genau eine Antwortalternative richtig ist – handelt es sich bei einer Aufgabe im *single-response*-Format um eine typische

Diskriminationsaufgabe: Da nur eine Antwort als richtig gilt, reicht es aus, diejenige zu wählen, die am ehesten dafür in Frage kommt – aus welchen Gründen auch immer. Vielleicht *wissen* die Prüflinge die richtige Antwort, finden diese unter den Antwortalternativen und wählen sie aus. Oder sie können eine oder auch mehrere der Antwortalternativen ausschließen, bis nur noch eine einzelne Antwortalternative übrig bleibt. Das tatsächliche Wissen der Prüflinge über die Korrektheit der einzelnen Antwortalternativen lässt sich mit dem *single-response*-Format deshalb so gut wie gar nicht abschätzen (Melzer, 2016). Die gewählte Antwortalternative war vielleicht nur die „am wenigsten falsch" erscheinende, oder unter lauter mehr oder weniger plausibel und richtig erscheinenden die plausibelste.

Für die Wissensdiagnose und die Frage, welche Art von Wissen sich mit Aufgaben im *single-response*-Format erfassen lässt, ist die Charakterisierung als Diskriminationsaufgabe von nicht zu unterschätzender Bedeutung und vermutlich den meisten Autoren und Autorinnen von Prüfungsaufgaben nicht hinlänglich bewusst. Es bedeutet nämlich, dass das zur richtigen Antwort nötige Wissen nicht nur von der korrekten Antwortalternative abhängt, sondern ganz entscheidend auch von den Distraktoren. Wenn etwa nach dem Geburtsjahr von Richard Wagner gefragt wird – die richtige Antwort ist 1813 –, dann macht es einen entscheidenden Unterschied, ob die Distraktoren 1810, 1811 und 1812 lauten oder: 1630, 1728 und 1845. Die Frage nach der Hauptstadt von Mosambik mit den Antwortalternativen Malabo, Maputo, Moroni und Maseru wird erheblich leichter, wenn die Antwortalternativen lauten: Kairo, Maputo, Nairobi und Tunis. Das liegt nicht an der Ratewahrscheinlichkeit – wer würfelt oder Münzen wirft, landet in beiden Fällen mit einer Wahrscheinlichkeit von ¼ bei der richtigen Antwort –, sondern es liegt daran, dass in den unterschiedlichen Versionen unterschiedliches Diskriminationswissen abgefragt wird.

6.2 Das *single-response*-Format in ILIAS

In ILIAS sind die folgende Aufgabentypen dem *single-response*-Format zuzuordnen:

- *Single Choice*
- *ImageMap* mit dem Antwortmodus „*Single Choice*"
- Lückentext vom Typ Auswahl-Lücke

6.2.1 *Single Choice*

Bei einer *Single-Choice*-Aufgabe wird den Prüflingen ein Aufgabentext und darunter eine Liste von Antwortalternativen präsentiert. Die Aufgabe der Prüflinge besteht darin, die im Sinne der Aufgabestellung zutreffende Antwortalternative – und nur diese – auszuwählen und den zugehörigen runden Button vor dieser Antwortalternative anzuklicken. ILIAS lässt nur die Auswahl einer einzigen Antwortalternative zu.

Dieser Aufgabentyp ist unter anderem in der Medizin weit verbreitet und wird dort auch als „Typ-A-Aufgabe" bezeichnet (Case & Swanson, 2002). Weiterhin ist dieser Aufgabentyp aus der populären Fernsehsendung „Wer wird Millionär" (Endemol, 2016) bekannt.

Informationen zum Erstellen einer *Single-Choice*-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 6.1 dargestellt.

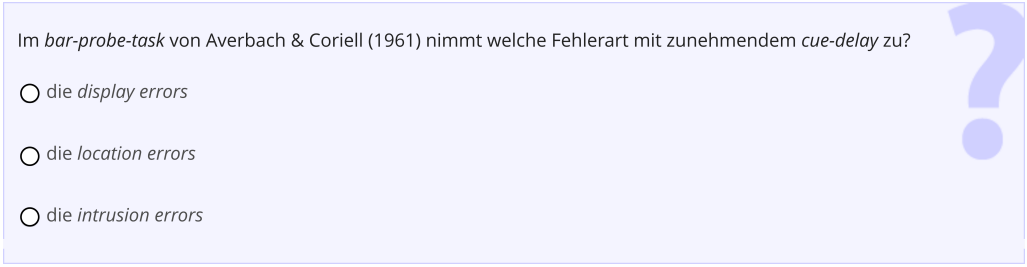


Abbildung 6.1. Beispiel einer *Single-Choice*-Aufgabe in ILIAS mit drei Antwortalternativen. Die richtige Lösung ist, nur die zweite Antwortalternative „die *location errors*" auszuwählen.

6.2.2 *ImageMap* mit dem Antwortmodus „*Single Choice*"

Bei einer *ImageMap*-Aufgabe mit dem Antwortmodus „*Single Choice*" wird den Prüflingen ein Aufgabentext und darunter ein Bildelement präsentiert. Bei dem Bildelement kann es sich zum Beispiel um ein Foto, eine Abbildung, eine Tabelle, einen Funktionsgraphen etc. handeln. Die Aufgabe der Prüflinge besteht darin, einen Bereich des Bildelements durch Anklicken zu markieren.

Dabei sind mehrere anklickbare Bereiche durch den Prüfenden vorgegeben, welche jedoch von Seiten des Systems nicht gesondert für die Prüflinge hervorgehoben werden. Die vorgegebenen Antwortalternativen können aber durch Bewegungen mit der Maus über das Gesamtbild

nach und nach erkannt werden, da sich der Mauscursor beim Überfahren von Antwortbereichen von einem Pfeil in eine Hand ändert. Es kann daher empfehlenswert sein, die als Antwortalternativen vorgesehenen Bereiche bereits beim Erstellen der Abbildung gesondert und für die Prüflinge erkennbar zu markieren.

Informationen zum Erstellen einer *ImageMap*-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 6.2 dargestellt.

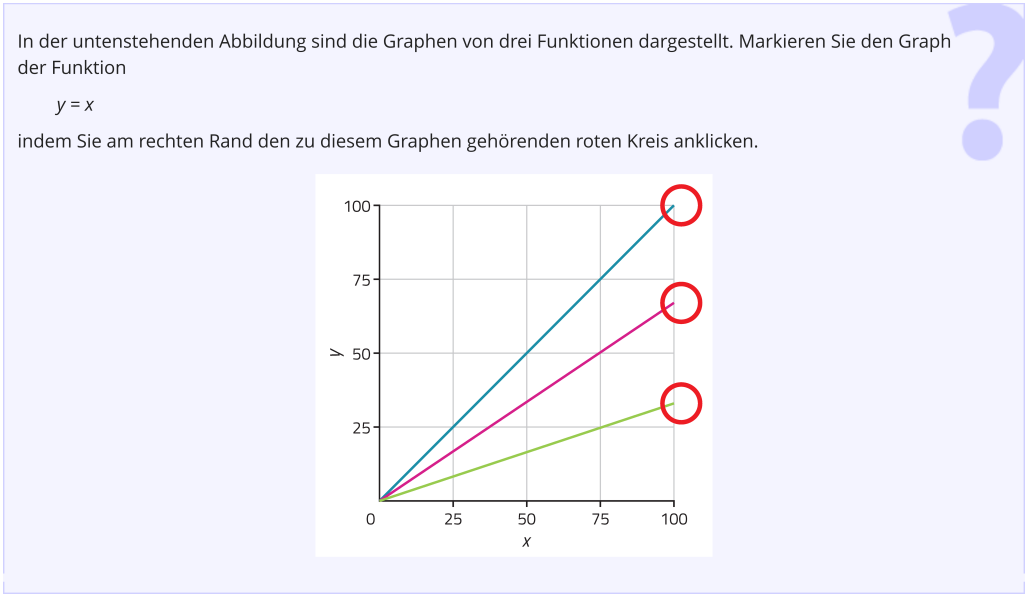


Abbildung 6.2. Beispiel einer *ImageMap*-Aufgabe mit dem Antwortmodus „*Single Choice*" in ILIAS mit drei Antwortalternativen. Die richtige Lösung ist, nur den roten Kreis rechts neben der blauen Linie anzuklicken.

6.2.3 Lückentext vom Typ „Auswahl-Lücke"

Bei einer Lückentext-Aufgabe wird den Prüflingen ein unvollständiger Text präsentiert. Die Aufgabe der Prüflinge beim Typ „Auswahl-Lücke" ist es, diesen Text an den vorgegebenen Stellen mit dem jeweils richtigen Begriff zu ergänzen. Dazu stehen in jeder Lücke mehrere hinterlegte Antwortalternativen in einem Dropdown-Menü zur Verfügung, von denen nur genau eine ausgewählt werden kann.

Informationen zum Erstellen einer Lückentext-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 6.3 dargestellt.

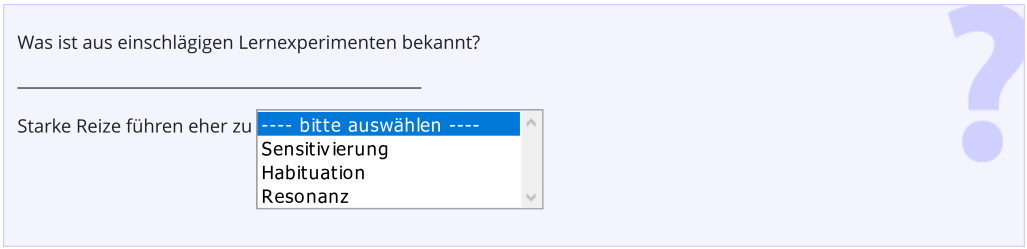


Abbildung 6.3. Beispiel einer Lückentext-Aufgabe vom Typ „Auswahl-Lücke" in ILIAS mit drei Antwortalternativen. Die richtige Lösung ist, den Listeneintrag „Sensitivierung" auszuwählen.

6.3 Parameter

6.3.1 Anzahl der Antwortalternativen *m* und Ratewahrscheinlichkeit *g*

Der wichtigste Parameter beim *single-response*-Format ist die Anzahl der verschiedenen Antwortalternativen, die bei einer Aufgabe vorgegeben sind. Er wird in diesem Handbuch mit m bezeichnet und hängt unmittelbar zusammen mit der Wahrscheinlichkeit dafür, bei einer rein zufälligen Auswahl einer Antwortalternative gerade die richtige zu treffen (Ratewahrscheinlichkeit; Fehler erster Art). Bei m-vielen Antwortalternativen, von denen genau eine richtig ist, beträgt die Ratewahrscheinlichkeit $g = 1/m$, da beim „blinden Raten" alle Antwortalternativen dieselbe Chance haben, gewählt zu werden.

6.3.2 Flüchtigkeitsfehler *f*

Die Wahrscheinlichkeit für einen Flüchtigkeitsfehler, also die Wahrscheinlichkeit dafür, dass ein Prüfling trotz sicheren Wissens eine falsche Antwort gibt, kann kaum zuverlässig geschätzt werden (Ünlü, 2006). Für eine möglichst realistische und faire Auswertung von Klausuren kann von Prüfenden ein plausibler Wert angesetzt werden, typischerweise ein Wert nahe null, z. B. $f = .05$ oder $f = .01$, je nachdem, wie viel Gewicht diesem Fehler zweiter Art beigemessen

wird. Natürlich kann für f auch der Wert null eingesetzt werden, so dass Flüchtigkeitsfehler unberücksichtigt bleiben.

6.3.3 Rateneigung *h*

Die Wahrscheinlichkeit dafür, dass ein Prüfling im Zweifel die Aufgabe nicht unbeantwortet lässt, sondern zufällig eine der vorgegebenen Antworten auswählt (vgl. Abschnitt 2.2), wird mit h bezeichnet. Da wir die Ratewahrscheinlichkeit bei der Auswertung berücksichtigen und deshalb die Prüflinge – aus Gründen, die in Kapitel 4 näher erläutert sind – ausdrücklich auffordern, jede Aufgabe zu beantworten und „im Zweifel zu raten", kann die Rateneigung als Konstante $h = 1$ angenommen werden und taucht deshalb in den Formeln dieses Abschnittes nicht mehr auf.

6.4 Scoringverfahren

Bei *single-response*-Aufgaben gibt es nur zwei „vernünftige" Scoringverfahren:

- Das Standardverfahren aus Abschnitt 3.1 mit:
  - 1 Punkt für die richtige Antwort
  - 0 Punkten in allen anderen Fällen (keine Antwort, falsche Antwort, mehr als eine Antwort, etc.)
- *formula scoring* aus Abschnitt 3.2.2 mit:
  - 1 Punkt für die richtige Antwort
  - $-1/(m-1)$ Punkte für eine falsche Antwort
  - 0 Punkten in allen anderen Fällen (keine Antwort, mehr als eine Antwort, etc.)

Wegen der juristischen Problematik von Minuspunkten ist das Standardverfahren die empfehlenswertere Variante, bei der die Ratewahrscheinlichkeit bei der Festsetzung der Notengrenzen berücksichtigt wird (s. Abschnitt 6.6). Wer dennoch lieber auf eine Ratekorrektur durch Minuspunkte setzt, darf aus den in Abschnitt 3.2.1 genannten Gründen nicht schematisch für Falschantworten einen ganzen Minuspunkt vergeben, sondern muss den Punktabzug nach der oben angegebenen Formel der Anzahl der Antwortalternativen anpassen: Bei $m = 3$ Antwortalternativen wird für Falschantworten $1/2$ Punkt abgezogen, bei $m = 4$ Antwortalternativen $1/3$ Punkt usw.

6.5 Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte

Der Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Zufallsvariablen X („erreichte Punkte in der Klausur") hängt vom gewählten Scoringverfahren ab (s. Kapitel 3).

Für das Standardscoring gilt nach Abschnitt 3.1:

$$\mathcal{E}(X) = n \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot p \cdot (1 - p)$$

mit:

$n :=$ Anzahl Aufgaben in der Klausur,

$p := p_W \cdot (1 - f - g) + g$,

$f :=$ Wahrscheinlichkeit für einen „Flüchtigkeitsfehler,"

$g := {}^1/_m$, Ratewahrscheinlichkeit bei m-vielen Antwortalternativen.

Für das *formula scoring* gilt nach Abschnitt 3.2.2 entsprechend:

$$\mathcal{E}(X) = n \cdot p$$

mit:

$n :=$ Anzahl Aufgaben in der Klausur,

$p := p_W \cdot (1 - {}^f/_{(1-g)})$,

$f :=$ Wahrscheinlichkeit für einen „Flüchtigkeitsfehler",

$g := {}^1/_m$, Ratewahrscheinlichkeit bei m-vielen Antwortalternativen.

In Abbildung 6.4 ist der Erwartungswert für eine Klausur bestehend aus 100 Aufgaben im *single-response*-Format in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt den Verlauf für das Standardscoring wieder, rechts ist der Verlauf für das *formula scoring* dargestellt. Auf diesen Funktionen beruht die Berechnung der Bestehens- und Notengrenzen im nächsten Abschnitt.

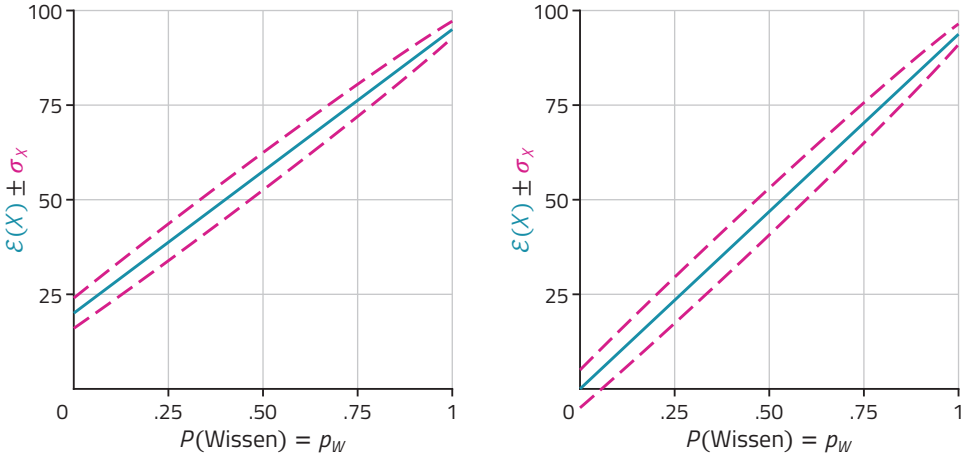


Abbildung 6.4. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim *formula scoring* (rechts). Beispiel für eine Prüfung bestehend aus $n = 100$ *single-response*-Aufgaben mit jeweils $m = 5$ Antwortalternativen ($g = 1/5$, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

Werden in einer Klausur Aufgaben im *single-response*-Format mit *unterschiedlicher* Anzahl von Antwortalternativen m verwendet, dann bildet man für jeden Wert von m eine eigene Aufgabengruppe. Die Erwartungswertfunktionen $\mathcal{E}(X) = F(p_W)$ werden zunächst getrennt für jede Gruppe erstellt (also z. B. für alle Aufgaben mit $m = 3$, alle Aufgaben mit $m = 4$ usw.). Die Gesamtfunktion ist schließlich die Summe aller einzelnen Funktionen. In Kapitel 12 wird dieses Verfahren zur Kombination beliebiger Aufgabenformate ausführlich erläutert.

6.6 Bestehens- und Notengrenzen

Wenn man von den Effekten des Flüchtigkeitsfehlers absieht, entsprechen die Erwartungswerte der erreichten Punkte beim *formula scoring* direkt dem Wissen der Probanden (s. Abbildung 6.4, rechts). Prüflinge, die 50% wissen ($p_W = .50$), haben einen Erwartungswert von 50% der maximalen Punkte, Prüflinge, die 95% wissen, erreichen im Durchschnitt 95% der maximalen Punkte usw. Man könnte die Bestehens- und Notengrenzen also unmittelbar in Prozentzahlen der maximal erreichbaren Punkte angeben, z. B.: bestanden bei mindestens

50%, Note 1.0 bei mindestens 95% usw.

Beim Standardscoring ist dies wegen der Ratewahrscheinlichkeit (Abbildung 6.4, links) nicht empfehlenswert. Wie in Kapitel 4 erläutert, müssen wir hier die ratekorrigierten Bestehens- und Notengrenzen, also die Erwartungswerte für $p_W = .50$, $p_W = .95$ usw., die wir mit $G_{korr}(q)$ bezeichnen, erst berechnen. Da es sich um lineare Funktionen handelt, ist das nicht weiter schwierig. Die allgemeine Formel dafür lautet:

$$G_{korr}(q) := n \cdot \big(g + q \cdot (1 - f - g)\big)$$

mit:

$n :=$ Anzahl Aufgaben in der Klausur,

$g :=$ $1/m$, Ratewahrscheinlichkeit bei m-vielen Antwortalternativen,

$f :=$ Wahrscheinlichkeit für einen „Flüchtigkeitsfehler,"

$q :=$ ursprüngliche Notengrenze als Anteil an der maximalen Punktzahl (z. B. $q = .50$ für das Bestehen bzw. $q = .95$ für die Note „sehr gut" etc.) und

$G_{korr}(q) :=$ ratekorrigierte Bestehens- bzw. Notengrenze als absoluter Punktwert (bei maximal n-vielen Punkten).

In Abbildung 6.5 und Tabelle 6.1 ist das beispielhaft für eine Klausur mit 100 *single-response*-Aufgaben dargestellt. Der Wert für f wurde in diesem Beispiel mit 0 angenommen. Bei $m = 5$ Antwortalternativen liegt die Bestehensgrenze für $p_W = .50$ bei 60% der maximal erreichbaren Punkte. Das entspricht im Übrigen genau der Grenze, die in der Approbationsordnung für Ärztliche Prüfungen (ÄApprO, 2013) bei den dort verwendeten Aufgaben („1 aus 5") festgelegt ist.

6.7 Zusammenfassung und Schlussfolgerung

Single response ist der am häufigsten verwendete Aufgabentyp. Diese Aufgaben sind aus den immer beliebter werdenden Quiz-Angeboten zu Unterhaltungszwecken gut bekannt. Vermutlich ist diese Vertrautheit der Hauptgrund für die große Verbreitung und die hohe Akzeptanz durch die Studierenden. Der Aufgabentyp ist, zumindest subjektiv, für Prüfende und Prüflin-

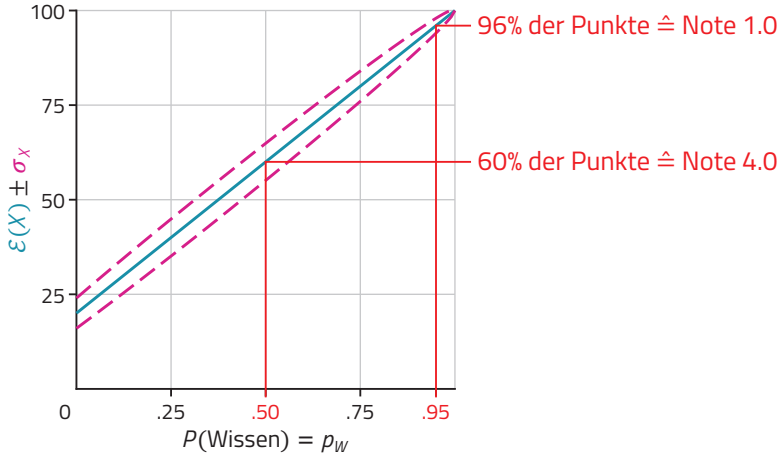


Abbildung 6.5. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiel für eine Prüfung mit $n = 100$ *single-response*-Aufgaben mit fünf Antwortalternativen ($g = .20$, $h = 1$, $f = 0$) und daher 100 erreichbaren Punkten. Die Punktwerte der Bestehensgrenze (Note 4.0) bei einem Wissen von $p_W = .50$ und der Bestnotengrenze (Note 1.0) bei einem Wissen von $p_W = .95$ sind rechts in rot angegeben.

ge gut durchschaubar. Die moderate Ratewahrscheinlichkeit erscheint Prüflingen „fair" und Prüfenden „noch akzeptabel". Als Vorteil wird gelegentlich gesehen, dass die Ratewahrscheinlichkeit leicht durch die Anzahl der Antwortalternativen modifiziert werden kann.

Aus einer inhaltlichen Perspektive sind sie für Prüfungen allerdings nur dann geeignet, wenn es darum geht, Wissen zu überprüfen, das in einer Diskriminanzleistung besteht. Das kann z. B. dann der Fall sein, wenn es zu einer richtigen Antwort mehrere häufig vorkommende oder naheliegende oder verbreitete Fehlkonzepte gibt.

Das Hauptproblem bei *single-response*-Aufgaben liegt darin, nicht-triviale Distraktoren zu finden. Wenn das schwierig ist, sollte man lieber einen anderen Aufgabentyp wählen. Distraktoren, die von vorneherein, also ohne relevantes Wissen, ausscheiden, führen zu einer schlecht kontrollierbaren Überschätzung des eigentlich interessierenden Wissens. Das Problem mit ungeeigneten Distraktoren wird zusätzlich dadurch verschärft, dass häufig eine hohe Anzahl von Distraktoren zur Verringerung der Ratewahrscheinlichkeit empfohlen wird und Prüfende deshalb dazu neigen, sich zu einer richtigen Antwort weitere, oft unsinnige oder schlecht passende, Distraktoren auszudenken, die zur Überprüfung des eigentlich interessierenden Wissens wenig beitragen.

Tabelle 6.1. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiele für Prüfungen aus $n = 100$ *single-response*-Aufgaben und daher 100 erreichbaren Punkten und verschiedene Anzahlen von Antwortalternativen m ($g = 1/m$, $h = 1$, $f = 0$).

p_W	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
					E(Punkte) in Prozent					
$m = 3$, $g \approx .33$	66.7	70.0	73.3	76.7	80.0	83.3	86.7	90.0	93.3	96.7
$m = 4$, $g = .25$	62.5	66.3	70.0	73.8	77.5	81.3	85.0	88.8	92.5	96.3
$m = 5$, $g = .20$	60.0	64.0	68.0	72.0	76.0	80.0	84.0	88.0	92.0	96.0
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Die generelle Empfehlung für *single-response*-Aufgaben lautet daher:

- *Single-response*-Aufgaben bieten sich an, wenn man zu einer Aufgabe inhaltlich sinnvoll eine korrekte und mehrere falsche Antwortalternativen formulieren kann und geprüft werden soll, ob Prüflinge die richtige Antwort von diesen falschen Antworten unterscheiden können.
- Es werden nur sinnvolle Distraktoren formuliert und keine weiteren „Pseudo-Distraktoren" zur Erhöhung der Distraktorenzahl.
- Da die Ratewahrscheinlichkeit ohnehin in Rechnung gestellt wird, wird darauf bei der Auswahl der Antwortalternativen keine Rücksicht genommen. Wenige sinnvolle Distraktoren sind auf jeden Fall besser als viele fragwürdige.
- Und vor allem: Es gibt viele andere Aufgabentypen und -formate, die vielleicht besser für das zu überprüfende Wissen geeignet sind.

7 Das *multiple-select*-Format

7.1 Charakteristik

Bei Aufgaben im *multiple-select*-Format (*MS*-Format) wird zu einer Aufgabe eine beliebige Anzahl *k* von Antwortalternativen vorgegeben, aus denen ein Prüfling *alle zutreffenden* auswählen soll. Aus einer Liste von Namen sind z. B. alle Nobelpreisträger auszuwählen, auf einer Landkarte alle Länder zu markieren, die zur EU gehören, oder für eine Menge von Medikamenten anzugeben, welche davon den Blutdruck senken.

Aufgaben im *multiple-select*-Format sehen äußerlich genauso aus wie Aufgaben im *single-response*-Format: Zu einer Aufgabe gibt es mehrere Antwortalternativen. Die Anforderung an die Prüflinge ist jetzt aber eine ganz andere. Da in der Regel nicht mitgeteilt wird, wie viele der Antwortalternativen tatsächlich das Auswahlkriterium erfüllen, müssen die Prüflinge für jede Antwortalternative getrennt entscheiden, ob sie diese für zutreffend halten oder nicht. Bei *k*-vielen Antwortalternativen sind also *k*-viele binäre Entscheidungen – ankreuzen oder nicht ankreuzen – zu treffen. Bei Aufgaben im *multiple-select*-Format handelt es sich damit um *Detektionsaufgaben* und nicht, wie beim *single-response*-Format, um Diskriminationsaufgaben (zur Unterscheidung von Diskrimination und Detektion in der Kognitionspsychologie und zu ihrer formalen Behandlung findet man in den klassischen Arbeiten von Luce, 1963 und Luce & Galanter, 1963 eine ausführliche Diskussion).

Für die Auswertung hat das Format von *multiple-select*-Aufgaben vor allem zwei Konsequenzen. Zum einen muss beim Erstellen der Antwortalternativen darauf geachtet werden, dass die einzelnen Entscheidungen unabhängig voneinander getroffen werden können. Alle statistischen Formulierungen in diesem Kapitel gehen davon aus, dass die *k*-vielen Entscheidungen stochastisch unabhängig sind. Zum anderen kann beim Scoring entweder jede einzelne Entscheidung für sich bewertet werden (Abschnitte 7.4.1 und 7.4.2) oder es kann das sogenannte

testlet scoring gewählt werden, bei dem Punkte für die Aufgabe insgesamt vergeben werden, z. B. nur dann, wenn alle Entscheidungen richtig sind (Abschnitt 7.4.3).

7.2 Das *multiple-select*-Format in ILIAS

In ILIAS sind die folgende Aufgabentypen dem *multiple-select*-Format zuzuordnen:

- *Multiple Choice*
- Fehlertext
- ImageMap mit dem Antwortmodus „*Multiple Choice*“

7.2.1 *Multiple Choice*

Bei einer *Multiple-Choice*-Aufgabe wird den Prüflingen zu einer Aufgabe eine Auswahl an vordefinierten Antwortalternativen präsentiert. Die Aufgabe der Prüflinge besteht darin, alle im Sinne der Aufgabenstellung zutreffenden Antwortalternativen durch Anklicken der dazugehörigen Kästchen auszuwählen. Dabei können keine, eine, mehrere oder auch alle angebotenen Antwortalternativen zutreffend sein.

Informationen zum Erstellen einer *Multiple-Choice*-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 7.1 dargestellt.

7.2.2 Fehlertext

Bei einer Fehlertext-Aufgabe wird den Prüflingen ein Text präsentiert, der fehlerhafte Wörter enthält. Die Aufgabe der Prüflinge ist es, diese Fehler zu erkennen und durch Anklicken der fehlerhaften Wörter im Text zu markieren, wobei jedes einzelne Wort anklickbar ist. Das Kriterium muss nicht die Orthographie sein, sondern kann vom Prüfenden durch die entsprechende Formulierung der Aufgabe vorgegeben werden, z. B. alle Verben oder alle Wörter, die groß geschrieben werden müssen, bis hin zu falschen Begriffen in einer Definition oder einem vorher auswendig zu lernenden Gedicht.

Informationen zum Erstellen einer Fehlertext-Aufgabe stehen im Wiki des @LLZ zur Verfügung. Ein einfaches Beispiel ist in Abbildung 7.2 dargestellt.

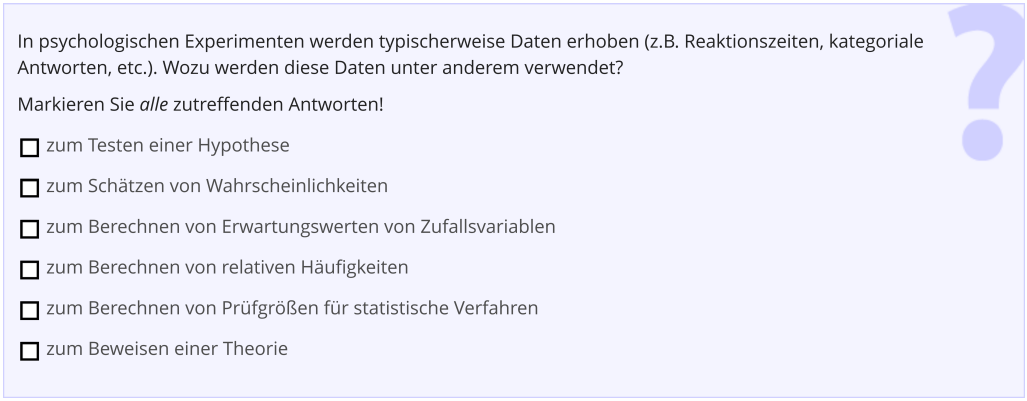


Abbildung 7.1. Beispiel einer *Multiple-Choice*-Aufgabe in ILIAS mit sechs Antwortalternativen. Die vollständig richtige Lösung ist, die Antwortalternativen eins, zwei, vier und fünf auszuwählen und keine weiteren.

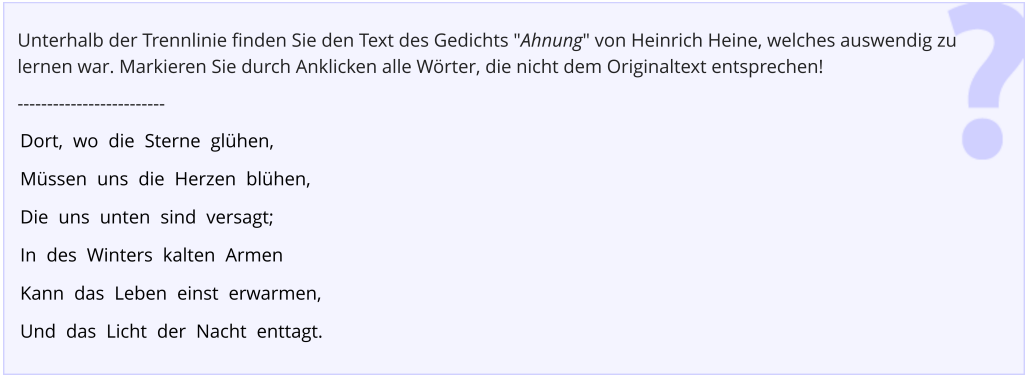


Abbildung 7.2. Beispiel einer Fehlertext-Aufgabe in ILIAS. Die vollständig richtige Lösung ist, die Wörter „Dort“ (richtig: „Oben“), „Herzen“ (richtig: „Freuden“), „Winters“ (richtig: „Todes“) und „einst“ (richtig: „erst“) zu markieren.

7.2.3 *ImageMap* mit dem Antwortmodus „*Multiple Choice*“

Bei einer *ImageMap*-Aufgabe mit dem Antwortmodus „*Multiple Choice*“ werden den Prüflingen ein Aufgabentext und darunter ein Bildelement präsentiert. Bei dem Bildelement kann es sich zum Beispiel um ein Foto, eine Abbildung, eine Tabelle, einen Funktionsgraphen etc. handeln. Die Aufgabe der Prüflinge besteht darin, alle Bereiche des Bildelements, die einem im Aufgabentext genannten Kriterium entsprechen, durch Anklicken zu markieren.

Dabei sind mehrere anklickbare Bereiche durch den Prüfenden vorgegeben, welche von Seiten des Systems nicht gesondert für die Prüflinge hervorgehoben werden. Allerdings können die Prüflinge sich die für die Antwort relevanten Bereiche erschließen, indem sie die Maus über das Bild bewegen und auf Veränderungen des Cursors achten. Es ist daher ggf. empfehlenswert, die relevanten Bereiche bereits beim Erstellen der Aufgabe im Quellbild zu markieren.

Informationen zum Erstellen einer *ImageMap*-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 7.3 dargestellt.

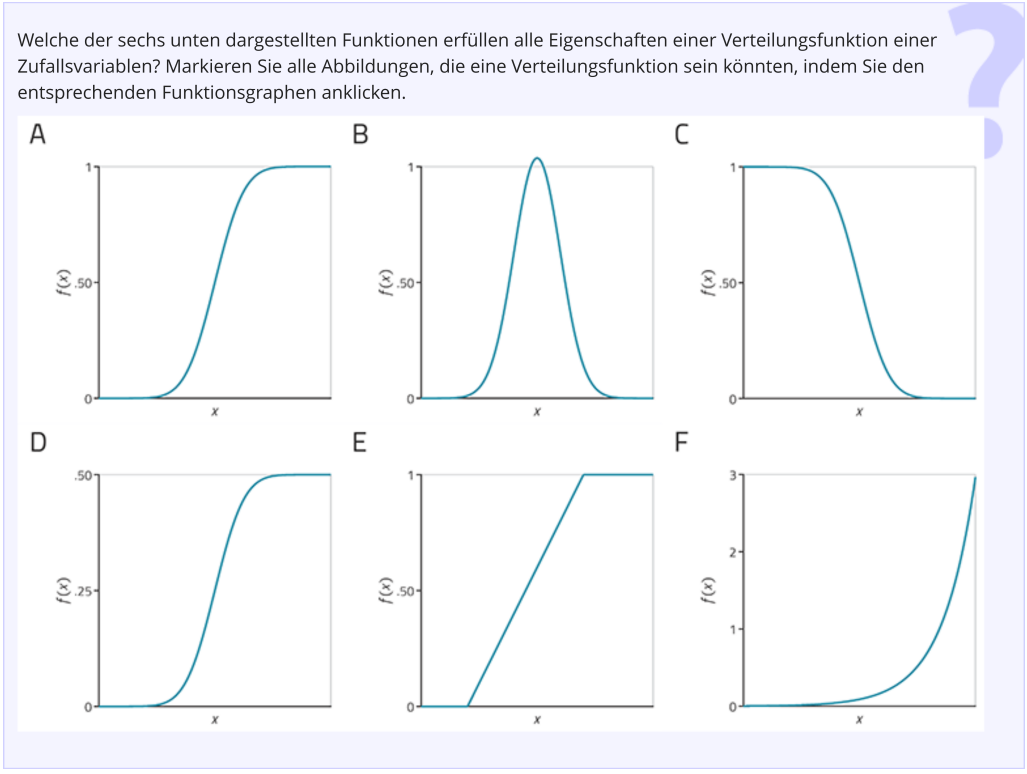


Abbildung 7.3. Beispiel einer *ImageMap*-Aufgabe mit dem Antwortmodus „*Multiple Choice*" in ILIAS mit sechs Antwortalternativen. Die vollständig richtige Lösung ist, die Funktionsgraphen A und E anzuklicken und keine weiteren.

7.3 Parameter

7.3.1 Anzahl der Antwortalternativen *k*

Die Anzahl der verschiedenen Antwortalternativen, die bei einer Aufgabe vorgegeben sind, wird beim *multiple-select*-Format mit *k* bezeichnet. Der Wert von *k* bestimmt die Anzahl der Entscheidungen, die zu treffen sind und darf nicht mit der Anzahl der Antwortmöglichkeiten verwechselt werden, auf die im nächsten Abschnitt 7.3.2 eingegangen wird.

7.3.2 Anzahl der Antwortmöglichkeiten *m*

Die Anzahl der Antwortmöglichkeiten wird mit *m* bezeichnet und beschreibt die Anzahl der Möglichkeiten, die den Prüflingen zur Beantwortung jeder einzelnen Antwortalternative zur Verfügung stehen. Im *multiple-select*-Format gibt es genau $m = 2$ Antwortmöglichkeiten, nämlich wählen (z. B. ankreuzen) bzw. nicht wählen (z. B. nicht ankreuzen). Für Prüfende ist es in diesem Format daher nicht erkennbar, ob eine Antwortalternative nicht angekreuzt wurde, weil sie als falsch eingeschätzt wurde oder weil sie ausgelassen wurde.

7.3.3 Ratewahrscheinlichkeit *g*

Die Anzahl der Antwortmöglichkeiten ist bei jeder Antwortalternative mit $m = 2$ – eben wählen bzw. nicht wählen – gegeben. Die Ratewahrscheinlichkeit beträgt demnach bei jeder Antwortalternative $g = 1/2$. Wegen der hohen Ratewahrscheinlichkeit beim *multiple-select*-Format sind hier geeignete Scoringverfahren besonders wichtig (s. Abschnitt 7.4).

7.3.4 Flüchtigkeitsfehler *f*

Die Wahrscheinlichkeit für einen Flüchtigkeitsfehler, also die Wahrscheinlichkeit dafür, dass ein Prüfling trotz sicheren Wissens eine falsche Antwort gibt, betrifft wieder jede einzelne Entscheidung. Wie bei allen Formaten kann vom Prüfenden ein plausibler Wert eingesetzt werden, typischerweise ein Wert nahe null, z. B. $f = .05$ oder $f = .01$, je nachdem, wie viel Gewicht diesem Fehler zweiter Art beigemessen wird. Natürlich kann für *f* auch der Wert null eingesetzt werden, so dass Flüchtigkeitsfehler unberücksichtigt bleiben.

7.3.5 Rateneigung *h*

Eine Besonderheit beim *multiple-select*-Format ist, dass es für die Prüflinge keine Möglichkeit gibt, eine Aufgabe nicht zu beantworten. Das Nicht-Auswählen einer Antwortalternative gilt immer als Entscheidung für das Kriterium „trifft nicht zu". Umso wichtiger ist deshalb die Anweisung an die Prüflinge, für jede Antwortalternative bewusst zu entscheiden, ob sie zutrifft oder nicht zutrifft und im Zweifel zu raten. Die Rateneigung in den Formeln dieses Abschnittes wird dementsprechend konstant auf den Wert $h = 1$ gesetzt und nicht mehr explizit als Variable geführt.

7.4 Scoringverfahren

Bei *multiple-select*-Aufgaben spielen Scoringverfahren aus zwei Gründen eine besonders wichtige Rolle. Zum einen ist hier die Ratewahrscheinlichkeit mit $g = .50$ besonders groß und muss zwingend berücksichtigt werden, zum anderen gibt es bei diesem Aufgabentyp die Möglichkeit zum *testlet scoring* mit einer Punktevergabe nur für die vollständig richtige Lösung der Aufgabe. Dementsprechend findet man bei diesem Aufgabenformat auch die vielfältigsten Formen der Punktvergabe. In Tabelle 7.1 sind die Scoringverfahren mit Bewertung jeder einzelnen Entscheidung zusammengefasst und werden im Folgenden genauer beschrieben. In Abschnitt 7.4.3 werden dann die Möglichkeiten zum *testlet scoring* erläutert.

7.4.1 Standardverfahren

Beim Standardverfahren (s. a. Abschnitt 3.1) zählen wir die Anzahl der richtigen Antworten. Als richtige Antwort gilt beim *multiple-select*-Format sowohl das Ankreuzen einer richtigen Antwortalternative als auch das Nicht-Ankreuzen einer falschen Antwortalternative. Wer bei k-vielen Antwortalternativen alles richtig macht, erhält k-viele Punkte. Die Anzahl der richtigen Antworten bzw. der richtigen Entscheidungen bezeichnen wir für dieses Scoring mit Y_{NR}, wobei *NR* für „*number right*" steht.

Die gelegentlich anzutreffende Variante, nur die Anzahl der richtigen Kreuze zu zählen, ist offensichtlich keine sinnvolle Scoringmethode und wurde deshalb nicht in die Übersicht in Tabelle 7.1 mit aufgenommen. Die Prüflinge könnten bei diesem Verfahren durch blindes

Tabelle 7.1. Übersicht über drei Scoringverfahren für *multiple-select*-Aufgaben mit Bewertung jeder einzelnen Entscheidung. Dargestellt ist für die einzelnen Scoringverfahren die Anzahl der vergebenen Punkte bei richtigen (R) bzw. falschen (F) Antwortalternativen, sofern Prüflinge diese ankreuzen (☒) bzw. nicht ankreuzen (☐).

Scoringverfahren	*formula scoring* (*FS*)			Standardverfahren (*number right*, *NR*)			*right minus wrong* (*RW*)		
Auszahlungsmatrix		☒	☐		☒	☐		☒	☐
	R	1	-1	R	1	0	R	1	0
	F	-1	1	F	0	1	F	-1	0
Summenscore	Y_{FS}			Y_{NR}			Y_{RW}		
	$Y_{FS} = 2 \cdot Y_{NR} - N$			$Y_{NR} = Y_{RW} + N_{falsch}$					

N := Gesamtanzahl der Antwortalternativen einer Prüfung
N_{falsch} := Anzahl der vorgegebenen falschen Antwortalternativen

Ankreuzen aller Antwortalternativen auch ohne jedes Wissen immer die maximale Punktzahl erreichen.

Tipp: Bei manchen Programmen zur automatisierten Punktvergabe ist es nur möglich, Punkte für positive Auswahlen, also angekreuzte Antwortalternativen zu vergeben, „kein Kreuz" wird dagegen immer mit null Punkten bewertet. In diesen Fällen lässt sich die Anzahl richtiger Entscheidungen sehr einfach berechnen durch:

$$Y_{NR} = Y_{RW} + N_{falsch}. \tag{7.1}$$

Dabei steht *RW* für „*right minus wrong*" und Y_{RW} gibt die Differenz zwischen der Anzahl der richtig gesetzten und der falsch gesetzten Kreuze an (s. a. Tabelle 7.1 und Abschnitt 7.4.2). N_{falsch} bezeichnet die Anzahl der vorgegebenen falschen Antwortalternativen.

7.4.2 Maluspunkte

Das naive Maluspunkteverfahren aus Abschnitt 3.2.1 – einen Punkt für eine richtige Antwort, einen Minuspunkt für eine falsche Antwort – entspricht wegen $g = 1/2$ gerade dem *formula scoring* aus Abschnitt 3.2.2. Die Summe der auf diese Weise bestimmten Punkte bezeichnen wir mit Y_{FS}.

Beim *RW*-Scoring (*right minus wrong*) werden dagegen werden nur die tatsächlich vorhandenen Kreuze betrachtet und bewertet, ob diese bei einer richtigen oder falschen Antwortalternative gesetzt wurden. Wird eine Alternative nicht angekreuzt, gibt es dafür keine Punkte. Bei diesem Scoringverfahren wird die Summe der erreichten Punkte mit Y_{RW} bezeichnet:

$$Y_{RW} := \text{Anzahl richtiger Kreuze} - \text{Anzahl falscher Kreuze.}$$

Y_{RW} ist allerdings *kein* empfehlenswertes Maß für das Wissen von Prüflingen. Es bringt alle Nachteile eines Maluspunkt-Scorings mit sich (juristische Probleme, negative Punktwerte, schlechte Interpretierbarkeit etc.) und es ist darüber hinaus direkt abhängig von der Anzahl der falschen Antwortalternativen (s. Gleichung 7.1). Sind *alle* Antwortalternativen einer Aufgabe falsch, dann ist der maximale Punktwert, also das bestmögliche Ergebnis für diese Aufgabe, null Punkte! Allerdings kann die Bestimmung von Y_{RW} als Zwischenschritt durchaus sinnvoll sein, weil man daraus mit Gleichung 7.1 den Punktwert bei Verwendung des Standardverfahrens Y_{NR} berechnen kann.

Das *formula scoring* dagegen ist ein ernstzunehmendes Scoringverfahren zur Ratekorrektur der Testergebnisse (vgl. Abschnitt 3.2.2). Beim *multiple-select*-Format ist Y_{FS} definiert durch:

$$Y_{FS} := \text{Anzahl richtiger Entscheidungen} - \text{Anzahl falscher Entscheidungen}$$

und es gilt:

$$Y_{FS} = 2 \cdot Y_{NR} - N, \tag{7.2}$$

wobei N für die Gesamtanzahl der Antwortalternativen einer Prüfung steht.

Gleichung 7.2 zeigt den linearen Zusammenhang zwischen Y_{FS} und Y_{NR} und gleichzeitig die Funktionsweise der Ratekorrektur. Während Y_{NR} Werte zwischen 0 und N annimmt, liegt der Wertbereich von Y_{FS} zwischen $-N$ und N.

7.4.3 *Testlet Scoring*

Beim *multiple-select*-Format sind innerhalb einer Aufgabe mehrere Entscheidungen zu treffen. Ein weit verbreitetes Scoringverfahren besteht darin, Punkte für eine Aufgabe erst dann zu vergeben, wenn alle Entscheidungen innerhalb dieser Aufgabe richtig getroffen wurden, wenn also alle zutreffenden Antwortalternativen – und nur diese – ausgewählt wurden. Diese

testlet-scoring-Variante (Wainer, Bradlow & Wang, 2007; Wainer & Kiely, 1987) ist als Alles-oder-nichts-Verfahren (AoN) bekannt. In schwächeren Varianten werden für geringe Fehler Teilpunkte vergeben, z. B. halbe Punkte, wenn alle Entscheidungen bis auf eine einzige richtig sind (vgl. K', z. B. Krebs, 2004).

Allgemeine Anmerkungen zum *testlet scoring* findet man in Abschnitt 3.3 dieses Handbuches. Einer genaueren Untersuchung der Auswirkungen des *testlet scoring* beim *multiple-select*-Format sind die Abschnitte 7.5.2 und 7.5.3 gewidmet.

7.5 Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte

Der Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Zufallsvariablen X („erreichte Punkte in der Klausur") hängt auch beim *multiple-select*-Format entscheidend vom gewählten Scoringverfahren ab (s. Kapitel 3). Die Darstellung dieses Zusammenhangs liefert wieder eine gute Grundlage für die Entscheidung, welches Scoringverfahren für die Bewertung einer Prüfung am besten geeignet ist.

Wir gehen dabei davon aus, dass eine Klausur mit n-vielen *multiple-select*-Aufgaben vorliegt. Zu jeder Aufgabe gibt es k-viele Antwortalternativen, von denen jeweils die zutreffenden anzukreuzen sind. Mit dieser Festlegung bezeichnen wir mit

$n :=$ Anzahl der Klausuraufgaben,

$k :=$ Anzahl der vorgegebenen Antwortalternativen pro Aufgabe,

$N := n \cdot k$, Anzahl der Entscheidungen in der Klausur,

$X :=$ Anzahl der erreichten Punkte in der Klausur.

7.5.1 Vergleich des Standardverfahrens mit Maluspunkt-Scoring

Für das Standardscoring gilt nach Abschnitt 3.1:

$$\mathcal{E}(X) = n \cdot k \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot k \cdot p \cdot (1 - p)$$

mit:

$$p := p_W \cdot \left(\frac{1}{2} - f\right) + \frac{1}{2}, \text{ wegen } g = \frac{1}{2},$$

$$f := \text{Wahrscheinlichkeit für einen Flüchtigkeitsfehler.}$$

Für das *formula scoring* gilt nach Abschnitt 3.2.2 entsprechend für jede Aufgabe i und jede Antwortalternative j die Bewertung:

$$Y_{ij} := \begin{cases} 1 & \text{falls die Entscheidung bei Aufgabe } i \text{ und Antwortalternative } j \text{ richtig ist } (A_r) \\ -1 & \text{falls die Entscheidung bei Aufgabe } i \text{ und Antwortalternative } j \text{ falsch ist } (A_f). \end{cases}$$

Wegen

$$P(A_r) = p_W \cdot \left(\frac{1}{2} - f\right) + \frac{1}{2} \quad \text{und}$$

$$P(A_f) = p_W \cdot \left(f - \frac{1}{2}\right) + \frac{1}{2}$$

gilt für den Erwartungswert von Y_{ij}:

$$\begin{aligned} \mathcal{E}(Y_{ij}) &= P(A_r) - P(A_f) \\ &= p_W \cdot (1 - 2f) \end{aligned}$$

und für den Erwartungswert im Gesamttest:

$$\begin{aligned} \mathcal{E}(X) &= n \cdot k \cdot \mathcal{E}(Y_{ij}) \\ &= p_W \cdot n \cdot k \cdot (1 - 2f). \end{aligned}$$

In analoger Weise berechnen wir die Standardabweichung von X. Für jede Einzelentscheidung gilt:

$$\begin{aligned} \text{var}(Y_{ij}) &= \mathcal{E}(Y_{ij}^2) - \mathcal{E}(Y_{ij})^2 \\ &= 1 - {p_W}^2 \cdot (1 - 2f)^2 \end{aligned}$$

und damit:

$$\text{var}(X) = n \cdot k \cdot \left(1 - {p_W}^2 \cdot (1 - 2f)^2\right).$$

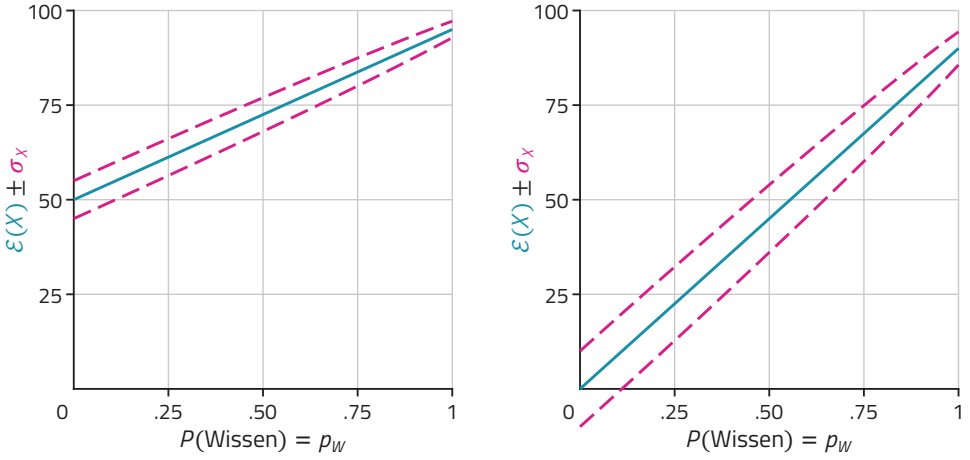


Abbildung 7.4. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim *formula scoring* (rechts). Beispiel für eine Prüfung bestehend aus $n = 20$ *multiple-select*-Aufgaben mit jeweils $k = 5$ Antwortalternativen ($g = 1/2$ für jede Antwortalternative, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

In Abbildung 7.4 ist der Erwartungswert für eine Klausur bestehend aus 20 Aufgaben im *multiple-select*-Format mit jeweils fünf Antwortalternativen in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt die Ergebnisse für das Standardscoring wieder, rechts ist der Verlauf für das *formula scoring* dargestellt. Die Ratekorrektur durch das *formula scoring* ist gut zu erkennen: Die statistisch zu erwartende Punktzahl entspricht ziemlich genau dem Wissen p_W, während man beim Standardscoring bereits ohne jedes Wissen im Durchschnitt auf 50% der Punkte kommt.

Bei Aufgaben im *multiple-select*-Format ist es im Übrigen sehr einfach möglich, Aufgaben mit unterschiedlich vielen Antwortalternativen anzubieten. Sowohl beim Standardverfahren als auch beim *formula scoring* wird jede einzelne Antwortalternative mit richtig oder falsch bewertet, so dass man bei der Auswertung alle Aufgaben dieses Formats zusammenfassen und in den angegebenen Formeln für N anstelle des Ausdrucks

$$N = n \cdot k$$

den allgemeineren Ausdruck

$$N := \sum_{i=1}^{n} k_i$$

verwenden kann.

7.5.2 Vergleich des Standardverfahrens mit dem Alles-oder-nichts-Verfahren

Für das Standardscoring gilt wie in Abschnitt 7.5.1:

$$\mathcal{E}(X) = n \cdot k \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot k \cdot p \cdot (1 - p)$$

mit:

$$p := p_W \cdot \left(\frac{1}{2} - f\right) + \frac{1}{2}, \text{ wegen } g = \frac{1}{2},$$

$f :=$ Wahrscheinlichkeit für einen Flüchtigkeitsfehler.

Für das Alles-oder-nichts-Scoring (AoN) derselben Klausur ist nach Abschnitt 3.3.1 die Summe der erreichten Punkte binomialverteilt mit den Parametern n und p^k. Allerdings liegt der Wertebereich der erreichbaren Punkte bei diesem Scoringverfahren zwischen 0 und n Punkten, da es pro Aufgabe höchstens einen Punkt gibt, während der Wertebereich beim Standardverfahren zwischen 0 und $N := n \cdot k$ Punkten liegt.

Für eine direkte Vergleichbarkeit der beiden Scoringmethoden ändern wir daher das AoN-Verfahren so, dass bei der vollständig richtigen Lösung einer Aufgabe k-viele Punkte vergeben werden, also genauso viele, wie im Fall des Standardscorings für eine vollständig richtig gelöste Aufgabe. Für nur teilweise richtig gelöste Aufgaben gibt es keine Punkte, d. h.: Für jede Aufgabe i ergibt sich der Punktwert X_i^* für diese Aufgabe als:

$$X_i^* := \begin{cases} k & \text{falls } Y_i = k \\ 0 & \text{sonst.} \end{cases}$$

Da $X_i^* = k \cdot X_i$ gilt, lässt sich der Gesamtpunktwert X^* für das AoN-Scoring einfach bestimmen

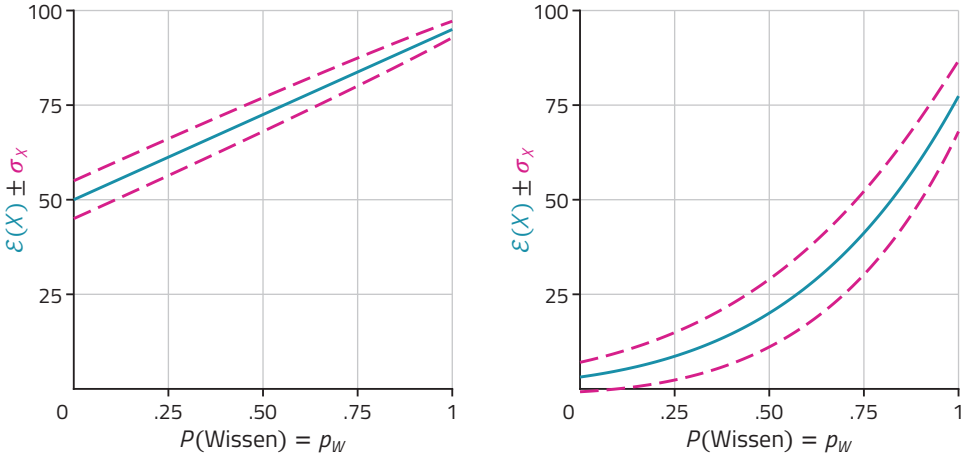


Abbildung 7.5. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim Alles-oder-nichts-Scoring (rechts). Beispiel für eine Prüfung bestehend aus $n = 20$ *multiple-select*-Aufgaben mit jeweils $k = 5$ Antwortalternativen ($g = 1/2$ für jede Antwortalternative, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

durch:

$$X^* := \sum_{i=1}^{n} X_i^*$$

und für dessen Erwartungswert und Varianz gilt:

$$\mathcal{E}(X^*) = k \cdot n \cdot p^k \quad \text{und}$$
$$\text{var}(X^*) = k^2 \cdot n \cdot p^k \cdot (1 - p^k).$$

In Abbildung 7.5 ist der Erwartungswert für eine Klausur bestehend aus 20 Aufgaben im *multiple-select*-Format mit jeweils fünf Antwortalternativen in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt die Ergebnisse für das Standardscoring wieder (und ist identisch mit der linken Grafik in Abbildung 7.4), rechts ist der Verlauf für das AoN-Scoring dargestellt.

Der Vergleich zeigt – noch deutlicher als in Abschnitt 3.3.1 – die Problematik des Alles-oder-nichts-Scorings. Die Ratewahrscheinlichkeit wird erfolgreich reduziert, der Preis dafür ist aber hoch: Der Erwartungswert für die erreichten Punkte in der Klausur steigt mit dem Wissen nur

sehr langsam an, bei einer Flüchtigkeitsfehlerwahrscheinlichkeit von $f = .05$ erreicht man selbst bei vollständigem Wissen $p_W = 1$ im Durchschnitt nur 77% der maximalen Punktzahl – das entspricht in vielen Prüfungsordnungen der Note 2.3. Dazu kommt, dass die Varianz hoch ist, insbesondere im Bereich zwischen $p_W = .50$ und $p_W = .75$, der bei Prüfungen am interessantesten ist, weil er den Großteil der Prüflinge umfasst.

7.5.3 Vergleich des Standardverfahrens mit dem K'-Scoring

Für das Standardscoring gilt weiterhin wie in Abschnitt 7.5.2:

$$\mathcal{E}(X) = n \cdot k \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot k \cdot p \cdot (1 - p)$$

mit:

$$p := p_W \cdot \left(\frac{1}{2} - f\right) + \frac{1}{2}, \text{ wegen } g = \frac{1}{2},$$
$$f := \text{Wahrscheinlichkeit für einen Flüchtigkeitsfehler.}$$

Um den im vorherigen Abschnitt 7.5.2 beschriebenen Auswirkungen des Alles-oder-nichts-Verfahrens entgegenzuwirken, werden in abgeschwächter Form oft für „fast richtige" Aufgabenlösungen Teilpunkte vergeben, z. B. beim K'-Scoring aus Abschnitt 3.3.2. Angewandt auf eine Klausur mit n-vielen Aufgaben und jeweils k-vielen Antwortalternativen werden beim K'-Scoring die Punkte zum Beispiel nach der Regel „volle Punktzahl, wenn alles richtig ist, halbe Punktzahl bei einem Fehler, keine Punkte bei mehr als einer falschen Entscheidung" zugewiesen. D. h.: für jede Aufgabe i ergibt sich der Punktwert X_i^* für diese Aufgabe als:

$$X_i^* := \begin{cases} k & \text{falls } Y_i = k \\ k/2 & \text{falls } Y_i = k - 1 \\ 0 & \text{sonst.} \end{cases}$$

Für den Erwartungswert von X_i^* gilt:

$$\mathcal{E}(X_i^*) = k \cdot \left(p^k + \frac{k}{2} \cdot p^{k-1} \cdot (1 - p)\right)$$

und für die Varianz:

$$\text{var}(X_i^*) = \mathcal{E}(X_i^{*2}) - \mathcal{E}(X_i^*)^2$$

mit:

$$\mathcal{E}(X_i^{*2}) = k^2 \cdot \left(p^k + \frac{k}{4} \cdot p^{k-1} \cdot (1-p) \right).$$

Das Klausurergebnis ist wieder der Summenscore X^*, der berechnet wird mittels:

$$X^* := \sum_{i=1}^{n} X_i^*.$$

Er nimmt Werte zwischen 0 und $n \cdot k$ Punkten an und es gilt:

$$\mathcal{E}(X^*) = n \cdot \mathcal{E}(X_i^*) \quad \text{und}$$
$$\text{var}(X^*) = n \cdot \text{var}(X_i^*).$$

In Abbildung 7.6 ist der Erwartungswert für eine Klausur bestehend aus 20 Aufgaben im *multiple-select*-Format mit jeweils fünf Antwortalternativen in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt die Ergebnisse für das Standardscoring wieder (und ist identisch mit der linken Grafik in Abbildung 7.4 und Abbildung 7.5), rechts ist der Verlauf für das K'-Scoring dargestellt.

Wie wir schon in Abschnitt 3.3.2 gesehen haben, ist der Verlauf des Erwartungswerts und die Größe der Varianz beim K'-Scoring nicht mehr ganz so dramatisch wie beim Alles-oder-nichts-Scoring. Wer alles weiß ($p_W = 1$) erreicht bei einer Flüchtigkeitsfehlerwahrscheinlichkeit von $f = .05$ im Durchschnitt immerhin 87.6% der maximalen Punktzahl – das entspricht in vielen Prüfungsordnungen der Note 1.7. Dennoch kann das nicht befriedigend sein, wir sprechen immerhin von jemandem, der alles weiß, also niemals rät. Auch perfekt vorbereitete Prüflinge erreichen im Schnitt nur die Note „gut" – mit einer Standardabweichung von über fünf Punkten, also einer ganzen Notenstufe. Die Varianz ist zwar insgesamt geringer als beim Alles-oder-nichts-Scoring, erreicht aber in der Spitze Werte von über acht Punkten.

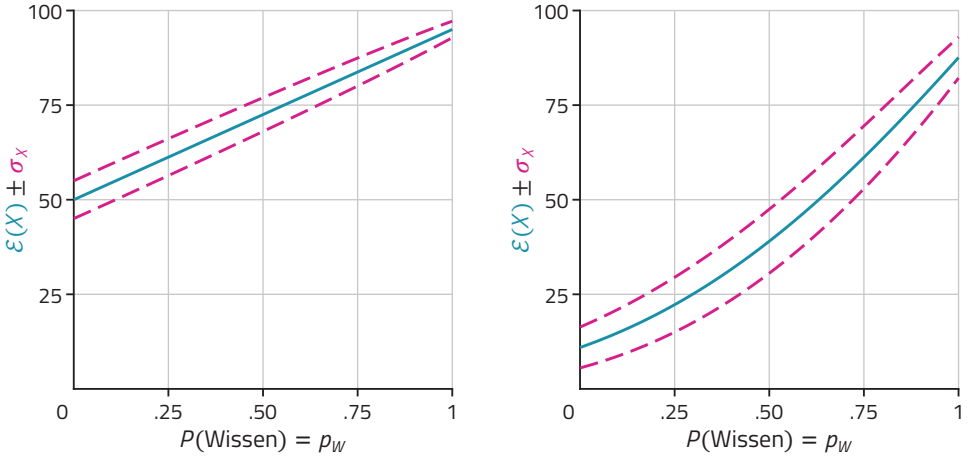


Abbildung 7.6. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim K'-Scoring (rechts). Beispiel für eine Prüfung bestehend aus $n = 20$ *multiple-select*-Aufgaben mit jeweils $k = 5$ Antwortalternativen ($g = 1/2$ für jede Antwortalternative, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

7.6 Bestehens- und Notengrenzen

Zusammenfassend zeigt Abschnitt 7.5 deutlich, dass auch beim *multiple-select*-Format die Erwartungswerte der erreichten Punkte nur beim *formula scoring* das Wissen der Probanden einigermaßen unverzerrt wiedergeben. Allerdings wird auch hier ein eventueller Flüchtigkeitsfehler nicht berücksichtigt. Beim Alles-oder-nichts-Scoring und beim K'-Scoring wird dieser Fehler sogar dramatisch verstärkt. Gegen diese beiden Verfahren spricht auch die Nichtlinearität der Erwartungswertfunktionen.

Am einfachsten und effektivsten ist auch hier die Verwendung des Standardscorings mit einer Anpassung der Bestehens- und Notengrenzen an die Ratewahrscheinlichkeit wie in Kapitel 4 beschrieben. Die Notengrenzen werden dabei durch das Wissen p_W definiert, z. B.: „bestanden hat, wer mindestens 50% weiß ($p_W \geq .50$)" oder „die Note 1.0 erhält, wer mindestens 95% weiß ($p_W \geq .95$)" oder wo immer Prüfende oder eine Prüfungsordnung die Grenzen setzen möchten.

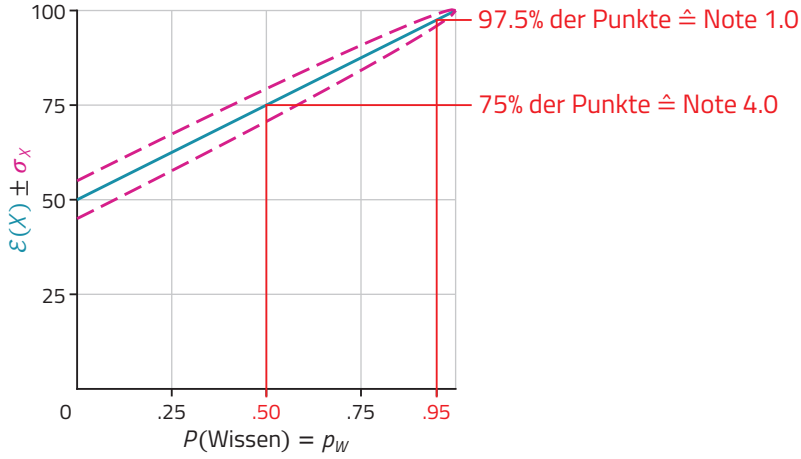


Abbildung 7.7. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiel für eine Prüfung aus $n = 25$ *multiple-select*-Aufgaben mit jeweils $k = 4$ Antwortalternativen und daher 100 erreichbaren Punkten ($g = .50$, $h = 1$, $f = 0$). Die Punktwerte der Bestehensgrenze (Note 4.0) bei einem Wissen von $p_W = .50$ und der Bestnotengrenze (Note 1.0) bei einem Wissen von $p_W = .95$ sind rechts in rot angegeben.

Wie viele Punkte dafür nötig sind, wird über die Erwartungswertfunktion bestimmt, die beim *multiple-select*-Format gegeben ist durch (vgl. Abschnitt 7.5.1):

$$\mathcal{E}(X) = n \cdot k \cdot \left(p_W \cdot \left(\frac{1}{2} - f \right) + \frac{1}{2} \right).$$

In Abbildung 7.7 und Tabelle 7.2 ist das beispielhaft für eine Klausur mit $n = 25$ Aufgaben im *multiple-select*-Format mit jeweils $k = 4$ Antwortalternativen wiedergegeben.

Rechnerisch lassen sich die Bestehens- und Notengrenzen leicht ermitteln. Bezeichnet man die unkorrigierten Grenzen mit G_{alt} und die neuen, ratekorrigierten Grenzen mit G_{neu} (jeweils in Prozentpunkten der maximal erreichbaren Punkte), dann gilt die einfache lineare Formel:

$$G_{neu} = G_{alt} \cdot \left(\frac{1}{2} - f \right) + 50.$$

Da beim *multiple-select*-Format die Ratewahrscheinlichkeit immer $g = 1/2$ beträgt und die Rateneigung h mit dem Wert 1 ebenfalls fest vorgegeben ist, bleibt als einzige Variable die vom Prüfenden festzulegende Flüchtigkeitsfehlertoleranz f. Bei einer ursprünglichen Bestehensgrenze von z. B. 50% der maximalen Punktzahl („bestanden hat, wer mindestens 50% des

Tabelle 7.2. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiel für eine Prüfung aus $n = 25$ *multiple-select*-Aufgaben mit jeweils $k = 4$ Antwortalternativen und daher 100 erreichbaren Punkten ($g = .50$, $h = 1$, $f = 0$).

p_W	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
					E(Punkte) in Prozent					
Einzelbewertung	75.0	77.5	80.0	82.5	85.0	87.5	90.0	92.5	95.0	97.5
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Stoffes beherrscht") und einer Flüchtigkeitsfehlertoleranz von $f = .05$, die den Prüflingen für „*careless errors*" zugestanden wird, ergibt sich daraus eine ratekorrigierte Bestehensgrenze von 72.5 % der maximalen Punkte.

7.7 Zusammenfassung und Schlussfolgerung

Aufgaben im *multiple-select*-Format erfordern ein absolutes Urteil über das Zutreffen eines Kriteriums. Die einzelnen Antwortalternativen müssen deshalb eindeutig in eine von zwei Kategorien klassifizierbar sein, wohingegen es beim *single-response*-Format ausreicht, wenn eine Antwortalternative das Kriterium *am besten* erfüllt. Die generelle Empfehlung für Aufgaben im *multiple-select*-Format lautet daher:

- Aufgaben im *multiple-select*-Format bieten sich an, wenn es um die Überprüfung einer Klassifikationsleistung geht. Aus einer Menge von vorgegebenen Antwortalternativen, z. B. Situationen, Objekten, Aussagen etc., sollen all diejenigen ausgewählt werden, die ein bestimmtes Kriterium erfüllen, z. B. alle richtigen, alle roten, alle wörtlichen Zitate aus dem Grundgesetz etc.
- Jede Antwortalternative muss eindeutig zu einer von zwei vorgegebenen Kategorien gehören, z. B. als „richtig" oder als „falsch" bewertet werden können.
- Anders als beim *single-response*-Format können beliebig viele richtige und falsche Antwortalternativen formuliert werden, da die Prüflinge für jede Antwortalternative eine Entscheidung treffen müssen.
- Die Ratewahrscheinlichkeit ist bei jeder Antwortalternative sehr hoch ($g = .50$). Sie muss deshalb bei der Festlegung der Bestehens- und Notengrenzen berücksichtigt werden.

- Aufgaben im *multiple-select*-Format sind außerordentlich flexibel und vielseitig einsetzbar. Die hohe Ratewahrscheinlichkeit ist weder ein Nachteil noch eine Einschränkung, da sie bei der Bewertung berücksichtigt wird.
- Aber wie immer gilt: Es gibt noch andere Aufgabentypen und -formate, die vielleicht für das in Frage stehende Wissen noch besser geeignet sind.

8 Das *multiple-true-false*-Format

8.1 Charakteristik

Bei Aufgaben im *multiple-true-false*-Format (*MTF*-Format) wird wie beim *multiple-select*-Format zu einer Aufgabe eine beliebige Anzahl *k* von Antwortalternativen vorgegeben, von denen einige richtig und die anderen falsch sind. Anders als beim *multiple-select*-Format sollen die Prüflinge hier aber bei jeder vorgegebenen Antwortalternative explizit ankreuzen, ob sie die Antwortalternative für richtig oder falsch halten. Für jede Antwortalternative sind daher zwei Antwortkästchen vorgesehen: eines für die Antwort „richtig" und eines für die Antwort „falsch" (s. a. Abbildung 8.2).

Die Aufgabe der Prüflinge ist dabei nicht auf eine richtig/falsch-Klassifikation beschränkt. Wie beim *multiple-select*-Format können auch beim *multiple-true-false*-Format die beiden Antwortkategorien beliebige dichotome Klassen darstellen, z. B.: ein Medikament senkt bzw. erhöht den Blutdruck; ein Tier ist ein bzw. ist kein Insekt; eine wässrige Lösung hat einen pH-Wert von größer bzw. kleiner sieben usw.

Die Anforderung an die Prüflinge ist beim *multiple-true-false*-Format sehr ähnlich zu der beim *multiple-select*-Format aus Kapitel 7, allerdings mit zwei kleinen, aber bedenkenswerten Unterschieden:

- Es gibt zu jeder vorgegebenen Antwortalternative nicht mehr zwei Antwortmöglichkeiten (ankreuzen bzw. nicht ankreuzen), sondern drei: „Kategorie A" ankreuzen, „Kategorie B" ankreuzen und kein Kreuz machen. Man kann bei diesem Format die Antwort also explizit verweigern bzw. im Zweifel *keine* Antwort geben.
- Die Anforderung, bei jeder Antwortalternative explizit eine Entscheidung treffen zu müssen, ist beim *multiple-true-false*-Format sehr viel deutlicher und transparenter als beim *multiple-select*-Format.

Als Konsequenz können wir eine Schlussfolgerung bereits vorwegnehmen: Wenn man die Wahl hat, sind Aufgaben im *multiple-true-false*-Format im Allgemeinen Aufgaben im *multiple-select*-Format vorzuziehen. Sie sind transparenter und eindeutiger sowohl in der Aufgabenstellung als auch in der Interpretation der Kreuze, die die Prüflinge setzen.

8.2 Das *multiple-true-false*-Format in ILIAS

In ILIAS sind Aufgaben im *multiple-true-false*-Format derzeit nur in einer sehr speziellen Variante, dem Aufgabentyp *Kprim Choice* verfügbar. Bei diesem Aufgabentyp ist sowohl die Anzahl der vorgegebenen Alternativen mit $k = 4$ als auch das Scoringverfahren fest vorgegeben: Punkte für die Beantwortung der Aufgabe bekommt nur, wer alle vier Antwortalternativen korrekt als „zutreffend“ oder „nicht zutreffend“ ankreuzt. Optional kann für Antworten mit drei richtigen Kreuzen die halbe Punktzahl vergeben werden (vgl. dazu die allgemeinen Ausführungen zum Aufgabentyp *Kprim* in Abschnitt 3.3).

Informationen zum Erstellen einer *Kprim-Choice*-Aufgabe stehen bei GoeEle@rn (2017) zur Verfügung. Ein einfaches Beispiel ist in Abbildung 8.1 dargestellt.

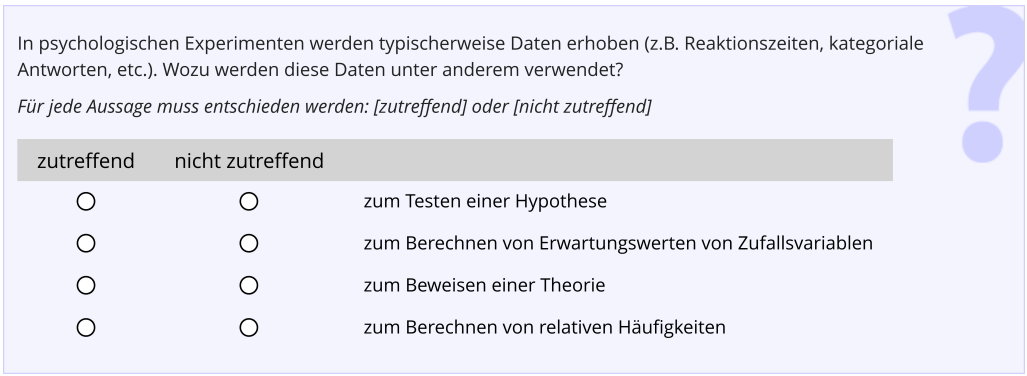


Abbildung 8.1. Beispiel einer *Kprim-Choice*-Aufgabe in ILIAS. Die vollständig richtige Lösung ist, bei den Antwortalternativen eins und vier „zutreffend“ anzukreuzen und bei den beiden anderen Antwortalternativen „nicht zutreffend“ anzukreuzen. Es handelt sich um eine auf $k = 4$ verkürzte Variante der Aufgabe, die auch im Kapitel zum *multiple-select*-Format in Abbildung 7.1 dargestellt ist.

Aufgaben im Allgemeinen *multiple-true-false*-Format mit beliebiger Anzahl von Alternativen und freier Wahl des Scoringverfahrens sind in ILIAS derzeit noch nicht vorgesehen. Ein ent-

sprechendes *Plugin* ist allerdings bereits beauftragt und wird wohl in einer späteren Version zur Verfügung gestellt. Als Übergangslösung kann man jede einzelne Antwortalternative als *Single-Choice*-Aufgabe mit zwei Antwortoptionen formulieren. Allerdings stehen dann nicht mehr alle in Abschnitt 8.4 besprochenen Scoringverfahren zur Verfügung (z. B. das Alles-oder-nichts-Verfahren) und bei Klausuren mit einer zufälligen Aufgabenreihenfolge können die zu einem Aufgabenstamm gehörenden Antwortalternativen auseinandergerissen werden.

Dennoch sei hier eine allgemeine Beispielaufgabe in Abbildung 8.2 dargestellt. Sie wurde mit EvaSys (Electric Paper Evaluationssysteme GmbH, 2017b) erstellt. Es handelt sich um die gleiche Aufgabe, die auch im Kapitel zum *multiple-select*-Format in Abbildung 7.1 verwendet wurde.

In psychologischen Experimenten werden typischerweise Daten erhoben (z.B. Reaktionszeiten, kategoriale Antworten, etc.). Wozu werden diese Daten unter anderem verwendet?

	falsch	richtig
zum Testen einer Hypothese	☐	☐
zum Schätzen von Wahrscheinlichkeiten	☐	☐
zum Berechnen von Erwartungswerten von Zufallsvariablen	☐	☐
zum Berechnen von relativen Häufigkeiten	☐	☐
zum Berechnen von Prüfgrößen für statistische Verfahren	☐	☐
zum Beweisen einer Theorie	☐	☐

Abbildung 8.2. Beispiel einer *multiple-true-false*-Aufgabe in EvaSys (Electric Paper Evaluationssysteme GmbH, 2017b) mit sechs Antwortalternativen. Die vollständig richtige Lösung ist, bei den Antwortalternativen eins, zwei, vier und fünf „richtig" anzukreuzen und bei den beiden anderen Antwortalternativen „falsch" anzukreuzen. Es handelt sich um die gleiche Aufgabe, die auch im Kapitel zum *multiple-select*-Format in Abbildung 7.1 dargestellt ist.

8.3 Parameter

8.3.1 Anzahl der Antwortalternativen *k*

Die Anzahl der verschiedenen Antwortalternativen, die bei einer Aufgabe vorgegeben sind, wird beim *multiple-true-false*-Format mit *k* bezeichnet. Der Wert von *k* bestimmt die Anzahl der

Entscheidungen, die zu treffen sind und darf nicht mit der Anzahl der Antwortmöglichkeiten verwechselt werden, auf die im nächsten Abschnitt 8.3.2 eingegangen wird.

8.3.2 Anzahl der Antwortmöglichkeiten *m*

Die Anzahl der Antwortmöglichkeiten wird mit *m* bezeichnet und beschreibt die Anzahl der Möglichkeiten, die den Prüflingen zur Beantwortung einer einzelnen Antwortalternative zur Verfügung stehen. Im *multiple-true-false*-Format gibt es wieder $m = 2$ Antwortmöglichkeiten, nämlich „Kategorie A" ankreuzen oder „Kategorie B" ankreuzen. Darüber hinaus können Prüflinge bei diesem Format auch keine Antwort geben, indem sie keine der Antwortalternativen ankreuzen oder eine ungültige Antwort geben, indem sie beide Antwortalternativen ankreuzen. Beim Scoring müssen diese Möglichkeiten zusätzlich berücksichtigt werden.

8.3.3 Ratewahrscheinlichkeit *g*

Da es nur $m = 2$ Antwortmöglichkeiten gibt, ist die Ratewahrscheinlichkeit wie beim *multiple-select*-Format für jede Antwortalternative mit $g = 1/2$ anzunehmen. Eine Berücksichtigung der Ratewahrscheinlichkeit bei der Bewertung einer Prüfungsleistung ist deshalb auch hier unerlässlich, entweder durch geeignete Scoringverfahren (Abschnitt 8.4) oder durch Anpassung der Bestehens- und Notengrenzen (Abschnitt 8.6).

8.3.4 Flüchtigkeitsfehler *f*

Die Wahrscheinlichkeit für einen Flüchtigkeitsfehler, also die Wahrscheinlichkeit dafür, dass ein Prüfling trotz sicheren Wissens eine falsche Antwort gibt, betrifft wieder jede einzelne Entscheidung. Wie bei allen Formaten kann vom Prüfenden ein plausibler Wert eingesetzt werden, typischerweise ein Wert nahe null, z. B. $f = .05$ oder $f = .01$, je nachdem, wie viel Gewicht diesem Fehler zweiter Art beigemessen wird. Natürlich kann für *f* auch der Wert null eingesetzt werden, so dass Flüchtigkeitsfehler unberücksichtigt bleiben.

8.3.5 Rateneigung *h*

Anders als bei Aufgaben im *multiple-select*-Format können sich Prüflinge beim *multiple-true-false*-Format aktiv dafür entscheiden, keine Antwort zu geben, wenn sie die Antwort nicht

wissen oder unsicher sind. Anderenfalls können sie raten. Die Rateneigung wird also eine Rolle spielen und in den Modellierungen zu berücksichtigen sein.

8.4 Scoringverfahren

Auch bei *multiple-true-false*-Aufgaben spielen Scoringverfahren wegen der hohen Ratewahrscheinlichkeit eine wichtige Rolle. Die Verhältnisse sind allerdings sehr übersichtlich, weil wir eine Aufgabe im *multiple-true-false*-Format als eine Zusammenfassung mehrerer *single-response*-Aufgaben mit jeweils zwei Antwortoptionen auffassen können. Für die Bewertung jeder einzelnen Entscheidung kommen deshalb nach Abschnitt 6.4 nur das Standardverfahren oder das *formula scoring* in Frage. Alternativ ist aber auch eine gemeinsame Bewertung aller Entscheidungen im Sinne des *testlet scoring* möglich.

8.4.1 Standardverfahren

Beim Standardverfahren (Abschnitt 3.1) zählen wir die Anzahl der richtigen Antworten zu den einzelnen Antwortalternativen. Als richtige Antwort zu einer Antwortalternative gilt im Falle des *multiple-true-false*-Formats, wie bei *single-response*-Aufgaben, wenn die richtige der beiden Antwortmöglichkeiten angekreuzt wird. Die Anzahl der richtigen Antworten bzw. der richtigen Entscheidungen bezeichnen wir für dieses Scoring mit Y_{NR}, wobei *NR* für „*number right*“ steht.

8.4.2 Maluspunkte

Das *formula scoring* aus Abschnitt 3.2.2 entspricht wegen der Ratewahrscheinlichkeit von $g = 1/2$ gerade dem naiven Maluspunkteverfahren aus Abschnitt 3.2.1, d. h.: Beim *multiple-true-false*-Format wird bei jeder Antwortalternative für eine richtige Antwort ein Punkt vergeben, für eine falsche Antwort ein Minuspunkt. Für ungültige oder fehlende Antworten werden keine Punkte vergeben. Die Summe der auf diese Weise bestimmten Punkte bezeichnen wir mit Y_{FS}.

Anders als beim *multiple-select*-Format gibt es beim *multiple-true-false*-Format aber keine feste numerische Beziehung zwischen Y_{NR} und Y_{FS}. Der Grund dafür ist, dass „keine Antwort“ beim Standardverfahren wie eine Falschantwort mit 0 Punkten bewertet wird, beim *formula scoring* diese beiden Fälle aber mit 0 Punkten bzw. -1 Punkt unterschiedlich bewertet werden.

8.4.3 *Testlet Scoring*

Beim *multiple-true-false*-Format sind wie beim *multiple-select*-Format für eine Aufgabe mehrere Entscheidungen zu treffen. Deshalb besteht auch hier die Möglichkeit, Punkte für eine Aufgabe erst dann zu vergeben, wenn alle (oder fast alle) Entscheidungen richtig getroffen wurden (vgl. Abschnitte 3.3 und 7.4.3).

8.5 Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte

Der Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Zufallsvariablen X („erreichte Punkte in der Klausur") hängt auch beim *multiple-true-false*-Format entscheidend vom gewählten Scoringverfahren ab (s. Kapitel 3). Die Darstellung dieses Zusammenhangs liefert wieder eine gute Grundlage für die Entscheidung, welches Scoringverfahren für die Bewertung einer Prüfung am besten geeignet ist.

Wir gehen dabei davon aus, dass eine Klausur mit n-vielen Aufgaben im *multiple-true-false*-Format vorliegt. Zu jeder Aufgabe gibt es k-viele Antwortalternativen, von denen jede mit „richtig" oder „falsch" bzw. „ja" oder „nein"; „trifft zu" oder „trifft nicht zu" etc. zu klassifizieren ist.

Mit dieser Festlegung bezeichnen wir wie in Kapitel 7 mit:

$n :=$ Anzahl der Klausuraufgaben,

$k :=$ Anzahl der vorgegebenen Antwortalternativen pro Aufgabe,

$N := n \cdot k$, Anzahl der Entscheidungen in der Klausur,

$X :=$ Anzahl der erreichten Punkte in der Klausur.

8.5.1 Vergleich des Standardverfahrens mit dem Maluspunkt-Scoring

Für das Standardscoring gilt nach Abschnitt 3.1:

$$\mathcal{E}(X) = n \cdot k \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot k \cdot p \cdot (1 - p)$$

mit:

$$p := p_W \cdot \left(1 - f - \frac{h}{2}\right) + \frac{h}{2},$$

$f :=$ Wahrscheinlichkeit für einen Flüchtigkeitsfehler.

Für die Ratewahrscheinlichkeit wird wieder ein fester Wert von $g = 1/2$ angenommen. Die Rateneigung wird dagegen zunächst noch als Variable h mitgeführt.

Für das *formula scoring* gilt nach Abschnitt 3.2.2 entsprechend für jede Aufgabe i und jede Antwortalternative j die Bewertung:

$$Y_{ij} := \begin{cases} 1 & \text{falls die Entscheidung bei Aufgabe } i \text{ und Antwortalternative } j \text{ richtig ist } (A_r) \\ -1 & \text{falls die Entscheidung bei Aufgabe } i \text{ und Antwortalternative } j \text{ falsch ist } (A_f) \\ 0 & \text{sonst.} \end{cases}$$

Wegen

$$P(A_r) = p_W \cdot \left(1 - f - \frac{h}{2}\right) + \frac{h}{2} \quad \text{und}$$
$$P(A_f) = p_W \cdot \left(f - \frac{h}{2}\right) + \frac{h}{2}$$

gilt für den Erwartungswert von Y_{ij}:

$$\begin{aligned} \mathcal{E}(Y_{ij}) &= P(A_r) - P(A_f) \\ &= p_W \cdot (1 - 2f) \end{aligned}$$

und für den Erwartungswert im Gesamttest:

$$\begin{aligned} \mathcal{E}(X) &= n \cdot k \cdot \mathcal{E}(Y_{ij}) \\ &= p_W \cdot n \cdot k \cdot (1 - 2f). \end{aligned}$$

Die Erwartunsgswerte beim *formula scoring* sind, wie wir in Abschnitt 3.2.2 gesehen haben, unabhängig von der Rateneigung h. Das gilt allerdings nicht für die Varianz. Wer rät, erhält mal einen Pluspunkt, mal einen Minuspunkt und im Durchschnitt null Punkte. Wer hingegen nicht rät und nichts ankreuzt, erhält auf jeden Fall null Punkte. Die Erwartungswerte sind deshalb für beide Strategien gleich, die Varianz ist aber beim Raten größer.

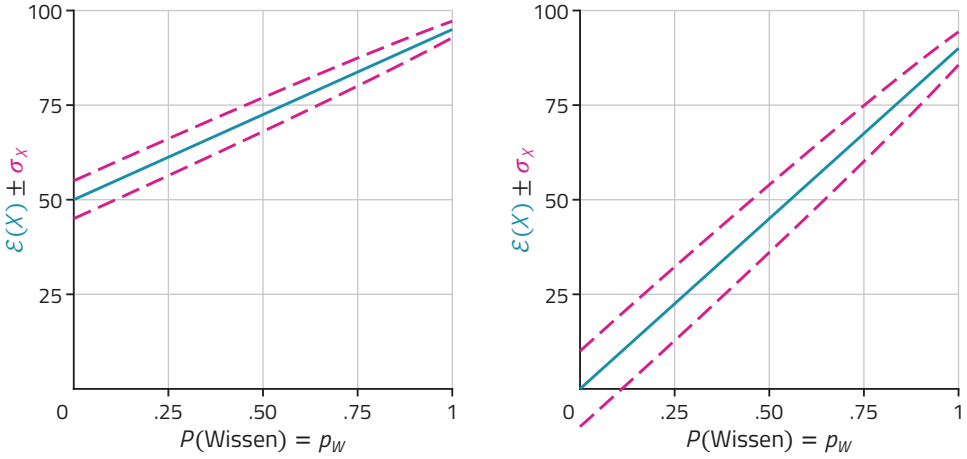


Abbildung 8.3. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim *formula scoring* (rechts). Beispiel für eine Prüfung bestehend aus $n = 20$ *multiple-true-false*-Aufgaben mit jeweils $k = 5$ Antwortalternativen ($g = 1/2$ für jede Antwortalternative, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

Wegen

$$\begin{aligned}\mathcal{E}(Y_{ij}^2) &= P(A_r) + P(A_f)\\ &= p_W \cdot (1-h) + h\end{aligned}$$

berechnen wir die Varianz für jede Einzelentscheidung mit:

$$\begin{aligned}\text{var}(Y_{ij}) &= \mathcal{E}(Y_{ij}^2) - \mathcal{E}(Y_{ij})^2\\ &= p_W \cdot (1-h) + h - {p_W}^2 \cdot (1-2f)^2\end{aligned}$$

und damit:

$$\text{var}(X) = n \cdot k \cdot \text{var}(Y_{ij}).$$

In Abbildung 8.3 ist der Erwartungswert für eine Klausur bestehend aus 20 Aufgaben im *multiple-true-false*-Format mit jeweils fünf Antwortalternativen in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt die Ergebnisse für das Standardscoring wieder, rechts ist

der Verlauf für das *formula scoring* dargestellt. Die Ratekorrektur durch das *formula scoring* ist gut zu erkennen: Die statistisch zu erwartende Punktzahl entspricht ziemlich genau dem Wissen p_W, während man beim Standardscoring bereits ohne jedes Wissen auf im Durchschnitt 50% der Punkte kommt. Für die Rateneigung wurde bei den Grafiken in der Abbildung ein Wert von $h = 1$ angenommen, da beim Standardscoring Raten die beste Strategie ist. Unter dieser Annahme stimmen die Kurven im Übrigen exakt mit denen aus Abbildung 7.4 für das *multiple-select*-Format überein. Das unterstreicht die formale Ähnlichkeit von *multiple-select*- und *multiple-true-false*-Aufgaben.

Bei Aufgaben im *multiple-true-false*-Format ist es im Übrigen ebenso wie beim *multiple-select*-Format sehr einfach möglich, Aufgaben mit unterschiedlich vielen Antwortalternativen anzubieten. Sowohl beim Standardverfahren als auch beim *formula scoring* wird jede einzelne Entscheidung bewertet, so dass man bei der Auswertung alle Aufgaben dieses Formats zusammenfassen und in den angegebenen Formeln für N anstelle des Ausdrucks

$$N = n \cdot k$$

den allgemeineren Ausdruck

$$N := \sum_{i=1}^{n} k_i$$

verwenden kann.

8.5.2 Vergleich des Standardverfahrens mit dem Alles-oder-nichts-Verfahren und dem K'-Scoring

Für das Standardscoring gilt wie in Abschnitt 8.5.1:

$$\mathcal{E}(X) = n \cdot k \cdot p \quad \text{und}$$
$$\text{var}(X) = n \cdot k \cdot p \cdot (1 - p)$$

mit:

$$p := p_W \cdot \left(1 - f - \frac{h}{2}\right) + \frac{h}{2},$$

$f :=$ Wahrscheinlichkeit für einen Flüchtigkeitsfehler.

Für das Alles-oder-nichts-Scoring (AoN) derselben Klausur ist nach Abschnitt 3.3.1 die Summe der erreichten Punkte binomialverteilt mit den Parametern n und p^k. Allerdings liegt der Wertebereich der erreichbaren Punkte bei diesem Scoringverfahren zwischen 0 und n Punkten, da es pro Aufgabe höchstens einen Punkt gibt, während der Wertebereich beim Standardverfahren zwischen 0 und $N := n \cdot k$ Punkten liegt.

Für eine direkte Vergleichbarkeit der beiden Scoringmethoden ändern wir daher das AoN-Verfahren so, dass bei der vollständig richtigen Lösung einer Aufgabe k-viele Punkte vergeben werden, also genauso viele wie im Fall des Standardscorings für eine vollständig richtig gelöste Aufgabe. Für nur teilweise richtig gelöste Aufgaben gibt es keine Punkte, d. h.: Für jede Aufgabe i ergibt sich der Punktwert X_i^* für diese Aufgabe als:

$$X_i^* := \begin{cases} k & \text{falls } Y_i = k \\ 0 & \text{sonst.} \end{cases}$$

Da $X_i^* = k \cdot X_i$ gilt, lässt sich der Gesamtpunktwert X^* für das AoN-Scoring bestimmen durch:

$$X^* := \sum_{i=1}^{n} X_i^*$$

und für dessen Erwartungswert und Varianz gilt:

$$\mathcal{E}(X^*) = k \cdot n \cdot p^k \quad \text{und}$$
$$\text{var}(X^*) = k^2 \cdot n \cdot p^k \cdot (1 - p^k).$$

In Abbildung 8.4 ist der Erwartungswert für eine Klausur bestehend aus 20 Aufgaben im *multiple-true-false*-Format mit jeweils fünf Antwortalternativen in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt die Ergebnisse für das Standardscoring wieder, rechts ist der Verlauf für das AoN-Scoring dargestellt. Beide Grafiken sind identisch mit denen aus Abbildung 7.5, da für die Rateneigung ein Wert von $h = 1$ angenommen wurde. Das ist gerade beim AoN-Scoring auch vernünftig: Wer unsicher ist und eine der Antwortalternativen nicht

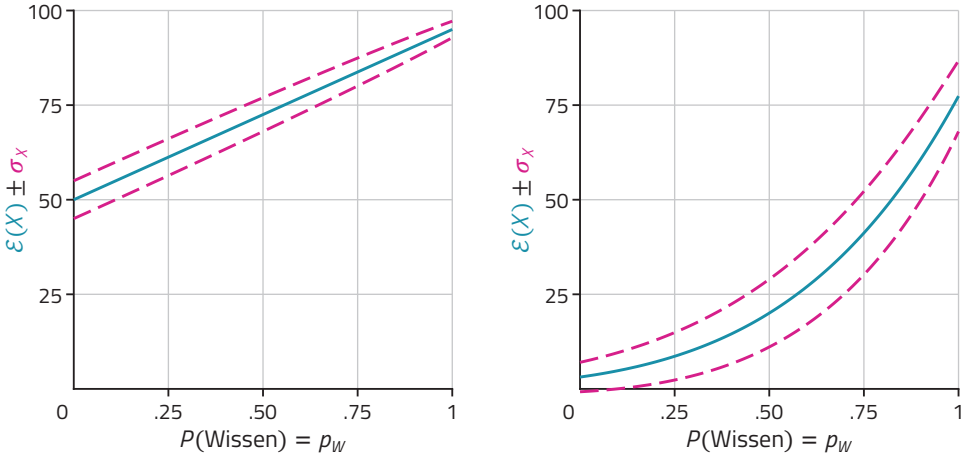


Abbildung 8.4. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim Alles-oder-nichts-Scoring (rechts). Beispiel für eine Prüfung bestehend aus $n = 20$ *multiple-true-false*-Aufgaben mit jeweils $k = 5$ Antwortalternativen ($g = 1/2$ für jede Antwortalternative, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

beantwortet, hat damit automatisch alle Punkte für diese Aufgabe verloren. Raten ist also beim AoN-Scoring geradezu obligatorisch, weil man nur gewinnen kann.

Der Vergleich zeigt wie in Abschnitt 7.5.2 die Problematik des Alles-oder-nichts-Scorings. Die Ratewahrscheinlichkeit wird erfolgreich reduziert, der Preis dafür ist aber hoch: Der Erwartungswert für die erreichten Punkte in der Klausur steigt mit dem Wissen nur sehr langsam an, bei einer Flüchtigkeitsfehlerwahrscheinlichkeit von $f = .05$ erreicht man selbst bei vollständigem Wissen $p_W = 1$ im Durchschnitt nur 77% der maximalen Punktzahl – das entspricht in vielen Prüfungsordnungen der Note 2.3. Dazu kommt, dass die Varianz hoch ist, insbesondere im Bereich zwischen $p_W = .50$ und $p_W = .75$, der bei Prüfungen am interessantesten ist, weil er den Großteil der Prüflinge umfasst.

Prüflinge, die – aus welchen Gründen auch immer – nicht raten, sondern bei Antwortalternativen, die sie nicht zuverlässig klassifizieren können, keine Antwort geben, vergeben mit dieser Strategie wertvolle Punkte. Bei beiden Scoringverfahren verringern sich mit abnehmender Rateneigung h die Erwartungswerte der Punkte, beim AoN-Scoring wirkt sich dies jedoch viel dramatischer aus als beim Standardscoring.

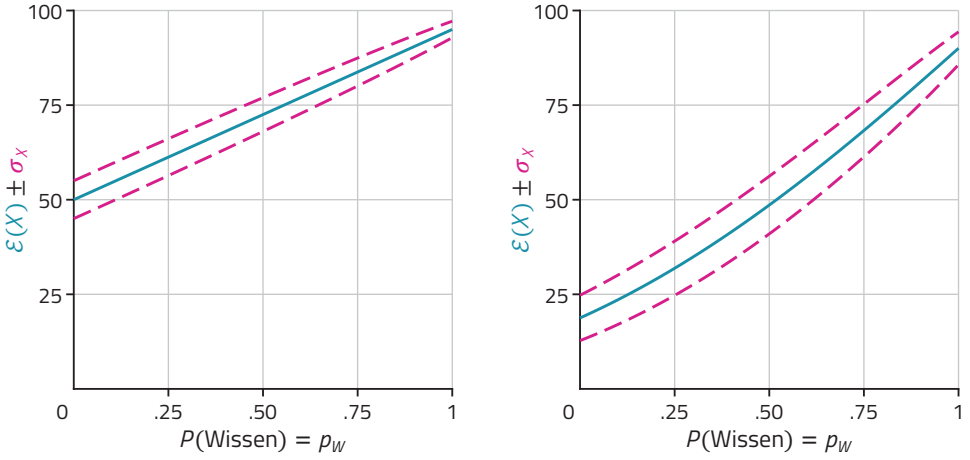


Abbildung 8.5. Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte (blaue Linie) beim Standardscoring (links) und beim K'-Scoring (rechts). Beispiel für eine Prüfung bestehend aus $n = 25$ *multiple-true-false*-Aufgaben mit jeweils $k = 4$ Antwortalternativen ($g = 1/2$ für jede Antwortalternative, $h = 1$). Für den Flüchtigkeitsfehler wurde ein Wert von $f = .05$ angenommen. Die pinkfarbenen, gestrichelten Linien geben $\pm$ eine Standardabweichung an. Zwischen diesen beiden Werten liegen für jedes Wissenslevel p_W etwa 68.4% der Prüflinge mit gleichem Wissen.

Für das K'-Scoring gilt im Übrigen beim *multiple-true-false*-Format dasselbe wie für das *multiple-select*-Format (s. Abschnitt 7.5.3), da die Berechnungen unter der Annahme $h = 1$ bei beiden Aufgabenformaten zum selben Ergebnis führen, wie wir am Beispiel des AoN-Scorings gesehen haben. In Abbildung 8.5 ist der Erwartungswert für eine Klausur bestehend aus 25 Aufgaben im *Kprim-Choice*-Format mit jeweils vier Antwortalternativen in Abhängigkeit vom Wissen p_W dargestellt. Die Grafik links gibt die Ergebnisse wieder, wie man sie für eine Auswertung nach dem Standardscoring erhalten würde, rechts ist der Verlauf für das (voreingestellte) K'-Scoring dargestellt.

Die Abbildung bestätigt, was wir bereits mehrfach festgestellt haben: Die Kurve nähert sich einer Geraden, ist aber noch erkennbar kurvilinear. Das K'-Scoring reduziert gegenüber dem Standardverfahren die Chance, Punkte durch bloßes Raten zu erreichen. Für $p_w = 0$ ist der Erwartungswert von knapp 20% aber immer noch erstaunlich hoch. Bei einer Flüchtigkeitsfehlerwahrscheinlichkeit von $f = .05$ ist ein Erwartungswert von nur etwa 90% für Prüflinge, die tatsächlich alles wissen ($p_w = 1$), ebenfalls nicht unproblematisch.

8.6 Bestehens- und Notengrenzen

Zusammenfassend zeigt Abschnitt 8.5 deutlich, dass auch beim *multiple-true-false*-Format die Erwartungswerte der erreichten Punkte nur beim *formula scoring* das Wissen der Probanden einigermaßen unverzerrt wiedergeben. Allerdings wird auch hier ein eventueller Flüchtigkeitsfehler nicht berücksichtigt. Beim Alles-oder-nichts-Scoring und beim K'-Scoring wird dieser Fehler sogar dramatisch verstärkt. Gegen diese beiden Verfahren spricht auch die Nichtlinearität der Erwartungswertfunktionen und die drastische Abhängigkeit von der Rateneigung.

Am einfachsten und nächstliegend ist deshalb auch hier die Verwendung des Standardscorings mit einer Anpassung der Bestehens- und Notengrenzen an die Ratewahrscheinlichkeit wie in Kapitel 4 beschrieben. Die Notengrenzen werden dabei durch das Wissen p_W definiert, z. B.: „bestanden hat, wer mindestens 50% weiß ($p_W \geq .50$)" oder „die Note 1.0 erhält, wer mindestens 95% weiß ($p_W \geq .95$)" oder wo immer Prüfende oder eine Prüfungsordnung die Grenzen setzen möchten.

Wie viele Punkte dafür nötig sind, wird über die Erwartungswertfunktion bestimmt, die beim *multiple-true-false*-Format gegeben ist durch (vgl. Abschnitt 8.5.1):

$$\mathcal{E}(X) = n \cdot k \cdot \left(p_W \cdot \left(1 - f - \frac{h}{2} \right) + \frac{h}{2} \right).$$

Für h wird man im Allgemeinen den Wert $h = 1$ einsetzen, da „im Zweifel raten" die beste Strategie ist.

In Abbildung 8.6 und Tabelle 8.1 ist das beispielhaft für eine Klausur mit $n = 25$ Aufgaben im *multiple-true-false*-Format mit jeweils $k = 4$ Antwortalternativen wiedergegeben.

Rechnerisch lassen sich die Bestehens- und Notengrenzen wieder leicht ermitteln. Bezeichnet man die unkorrigierten Grenzen mit G_{alt} und die neuen, ratekorrigierten Grenzen mit G_{neu} (jeweils in Prozentpunkten der maximal erreichbaren Punkte), dann gilt die einfache lineare Formel:

$$G_{neu} = G_{alt} \cdot \left(1 - f - \frac{h}{2} \right) + 50 \cdot h.$$

Die Ratewahrscheinlichkeit ist in dieser Gleichung mit $g = 1/2$ fest vorgegeben. Die Rateneigung h wird aber ebenso wie die vom Prüfenden festzulegende Flüchtigkeitsfehlertoleranz f als Variable explizit mit aufgeführt, um deren Einfluss deutlich zu machen. Bei einer ursprünglichen

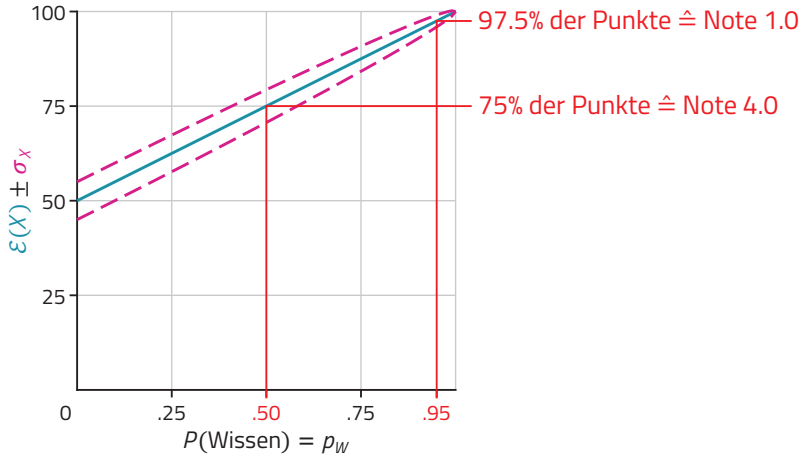


Abbildung 8.6. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiel für eine Prüfung aus $n = 25$ *multiple-true-false*-Aufgaben mit jeweils $k = 4$ Antwortalternativen und daher 100 erreichbaren Punkten ($g = .50$, $h = 1$, $f = 0$). Die Punktwerte der Bestehensgrenze (Note 4.0) bei einem Wissen von $p_W = .50$ und der Bestnotengrenze (Note 1.0) bei einem Wissen von $p_W = .95$ sind rechts in rot angegeben.

Bestehensgrenze von z. B. 50% der maximalen Punktzahl („bestanden hat, wer mindestens 50% des Stoffes beherrscht"), einer Rateneigung von $h = 1$ und einer Flüchtigkeitsfehlertoleranz von $f = .05$, die den Prüflingen für „*careless errors*" zugestanden wird, ergibt sich daraus eine ratekorrigierte Bestehensgrenze von 72.5% der maximalen Punkte.

Liegt die Rateneigung dagegen bei $h = 0$, was gleichbedeutend ist mit „niemals raten", dann müsste die neue Bestehensgrenze bei 47.5% der maximalen Punktzahl liegen und damit unter der ursprünglichen Grenze. Der Grund dafür ist, dass in diesem Fall nur der Fehler zweiter Art korrigiert wird, aber nicht die Ratewahrscheinlichkeit, da Raten mit der Annahme $h = 0$ ausgeschlossen wurde.

8.7 Zusammenfassung und Schlussfolgerung

Aufgaben im *multiple-true-false*-Format sind von den kognitiven Anforderungen her sehr ähnlich zu Aufgaben im *multiple-select*-Format. Der Charakter der Einzelentscheidung für jede einzelne Antwortalternative ist aber für den Prüfling deutlicher erkennbar und es gibt darüber hinaus die Möglichkeit, eine Antwortalternative bewusst nicht zu beantworten. Daher ist in

Tabelle 8.1. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiel für eine Prüfung aus $n = 25$ *multiple-true-false*-Aufgaben mit jeweils $k = 4$ Antwortalternativen und daher 100 erreichbaren Punkten ($g = .50$, $h = 1$, $f = 0$).

p_W	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
					E(Punkte) in Prozent					
Einzelbewertung	75.0	77.5	80.0	82.5	85.0	87.5	90.0	92.5	95.0	97.5
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Fällen, in denen beide Aufgabenformate möglich sind, das *multiple-true-false*-Format in der Regel zu bevorzugen. Abgesehen davon teilen sich diese beiden Aufgabenformate dieselben Vor- und Nachteile (s. Abschnitt 7.7).

9 Aufgaben mit offenem Format

9.1 Charakteristik

Die Bezeichnung „offene Aufgaben“ umfasst alle Aufgabenformate, bei denen – im Gegensatz zu den geschlossenen Aufgabenformaten – keine Antwortalternativen durch den Prüfenden vorgegeben sind. Diese Aufgaben gehören demnach nicht zu den Antwort-Wahl-Aufgaben und unterliegen daher nicht den strengen juristischen Anforderungen, die an das Antwort-Wahl-Format gestellt werden. Meist werden offene Aufgaben nicht gesondert in Prüfungs- oder Studienordnungen erwähnt bzw. werden keine speziellen Regeln zu deren Verwendung oder Bewertung aufgestellt. Da offene Aufgabenformate in ILIAS (und anderen Plattformen zum Erstellen von Prüfungsaufgaben) ausdrücklich vorgesehen sind und es durchaus sinnvoll ist, Klausuren mit Aufgaben im Antwort-Wahl-Verfahren gemischt mit Aufgaben im offenen Format zu erstellen, soll dieser Aufgabentyp hier kurz mit behandelt werden.

Hinsichtlich der Anforderungen, die offene Aufgaben an das Wissen von Prüflingen stellen, unterscheiden sich diese grundlegend von Antwort-Wahl-Aufgaben: Während die Prüflinge bei letzteren nur eine oder mehrere richtige Antworten aus einer größeren Menge von vorgegebenen Alternativen auswählen, müssen bei ersteren alle Antworten von den Prüflingen selbst produziert werden. Anstelle von *recognition*, also dem Wiedererkennen richtiger Antworten bei geschlossenen Aufgaben, erfordern offene Aufgaben *recall*, also den freien Abruf von Wissen.

Das freie Reproduzieren einer Antwort ist im Allgemeinen schwieriger als das Wiedererkennen einer vorgegebenen richtigen Antwort. Das entspricht nicht nur unserer Alltagserfahrung, sondern wurde auch in zahlreichen Studien nachgewiesen. Einen ersten Überblick dazu findet man z. B. bei Lieberman (2011). Im Vergleich zu geschlossenen Aufgaben sind also beim Einsatz von offenen Aufgaben deutlich geringere Leistungen zu erwarten. Darüber hinaus sind die Möglichkeiten zum erfolgreichen Erraten der richtigen Lösungen stark eingeschränkt.

9.2 Das offene Aufgabenformat in ILIAS

In ILIAS sind die folgende Aufgabentypen dem offenen Aufgabenformat zuzuordnen:

- Freitext
- Text-Teilmenge
- Lückentext (Numerische Lücke)
- Lückentext (Textlücke)
- *Long-Menu*-Frage
- Numerische Frage
- Formelfrage
- *JSME*-Frage

9.2.1 Freitext

Bei einer Freitext-Aufgabe wird eine offen formulierte Aufgabe gestellt oder ein Arbeitsauftrag erteilt. Die Aufgabe der Prüflinge besteht darin, selbstständig eine entsprechende Antwort in dem vorgegebenen Textfeld zu verfassen. Die Anzahl der maximal einzugebenden Zeichen kann begrenzt werden. Es kommen somit sowohl einzelne kurze Begriffe als auch längere Erläuterungen bzw. Definitionen oder auch ganze Aufsätze als mögliche Aufgaben in Frage.

ILIAS bietet bei diesem Aufgabentyp die Möglichkeit einer automatischen – allerdings sehr oberflächlichen – Bewertung. Dazu wird der durch die Prüflinge eingegebene Text auf das Vorkommen bestimmter, vorher festgelegter Begriffe überprüft. Eine darüber hinausgehende Bewertung, z. B. darauf, ob die Begriffe überhaupt in einem sinnvollen Zusammenhang stehen, ist nicht möglich. Die Vergabe von Punkten muss deshalb in der Regel manuell erfolgen.

Informationen zum Erstellen einer Freitext-Aufgabe stehen in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 9.1 dargestellt.

9.2.2 Text-Teilmenge

Bei einer Text-Teilmengen-Aufgabe besteht die Aufgabe der Prüflinge darin, eine vorgegebene Anzahl von Begriffen zu produzieren. Diese sind ähnlich wie beim Lückentext in vorgegebene freie Felder einzutragen.

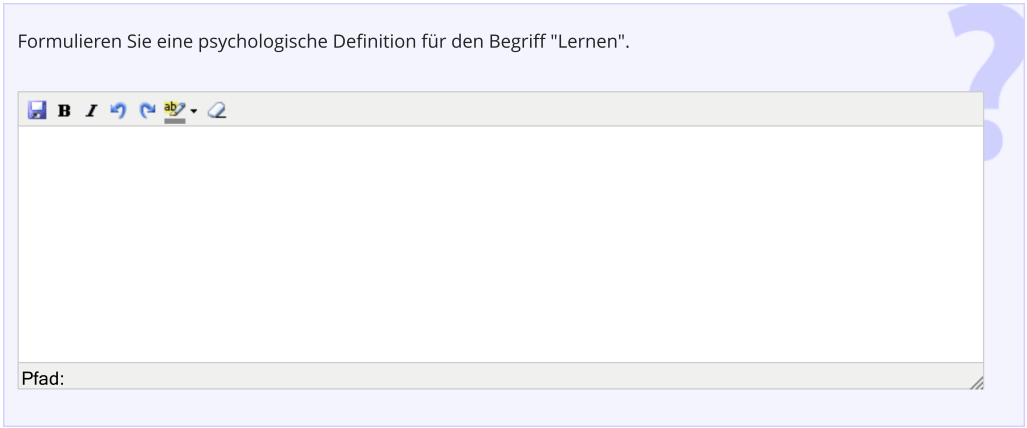


Abbildung 9.1. Beispiel einer Freitext-Aufgabe in ILIAS. Die richtige Antwort muss die Begriffe „Prozess“, „relativ überdauernd“, „Veränderung des Verhaltenspotenzials“ und „Erfahrung“ beinhalten.

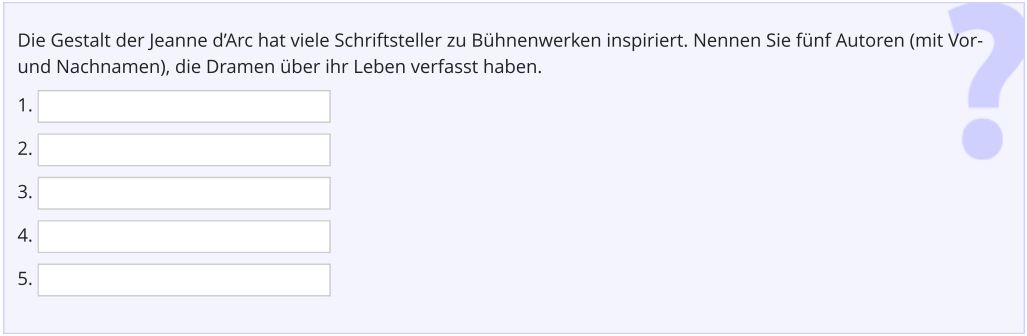


Abbildung 9.2. Beispiel einer Text-Teilmengen-Aufgabe in ILIAS. Zu den richtigen Antworten gehören z. B. William Shakespeare, Friedrich Schiller, Berthold Brecht, Jean Anouilh, Felix Mitterer, George Bernhard Shaw, etc.

ILIAS bietet bei diesem Aufgabentyp die Möglichkeit einer automatischen Bewertung. Auch hier werden die durch die Prüflinge eingegebenen Texte mit einer Liste von vorher festgelegten Begriffen verglichen. Dabei kann es jedoch vorkommen, dass die festgelegte Begriffsliste unvollständig ist und daher richtige Antworten unberücksichtigt bleiben.

Informationen zum Erstellen einer Text-Teilmengen-Aufgabe stehen in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 9.2 dargestellt.

9.2.3 Lückentext (Numerische Lücke)

Bei einer Lückentext-Aufgabe wird den Prüflingen ein unvollständiger Text präsentiert. Die Aufgabe der Prüflinge beim Typ „numerische Lücke“ ist es, den Text an den vorgegebenen Stellen mit den richtigen Zahlen zu ergänzen. Im Gegensatz zu numerischen Fragen (s. Abschnitt 9.2.6) können dabei mehrere Zahlen in einer Aufgabe abgefragt werden.

ILIAS bietet bei diesem Aufgabentyp die Möglichkeit einer automatischen Bewertung. Dazu wird die durch die Prüflinge in jeder Lücke eingegebene Zahl mit einer für diese Lücke vorher festgelegten richtigen Antwort verglichen. Es gibt die Möglichkeit, Intervalle festzulegen, in denen sich Lösungen befinden dürfen.

Informationen zum Erstellen einer Lückentext-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 9.3 dargestellt.

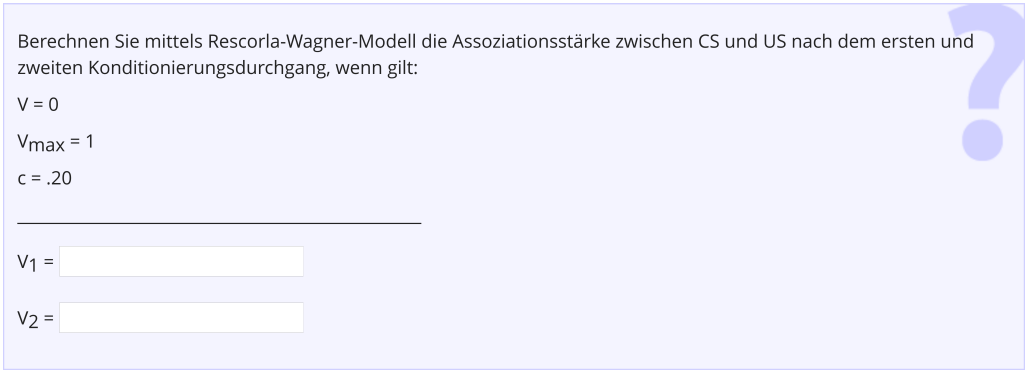


Abbildung 9.3. Beispiel einer Lückentext-Aufgabe vom Typ „numerische Lücke“ in ILIAS. Die richtigen Antworten sind „$V_1 = .20$“ bzw. „$V_2 = .36$“.

9.2.4 Lückentext (Textlücke)

Bei einer Lückentext-Aufgabe wird den Prüflingen ein unvollständiger Text präsentiert. Die Aufgabe der Prüflinge beim Typ „Textlücke“ ist es, diesen Text an den vorgegebenen Stellen mit den jeweils passenden Begriffen zu ergänzen.

ILIAS bietet bei diesem Aufgabentyp die Möglichkeit einer automatischen Bewertung. Dazu wird der durch die Prüflinge in jeder Lücke eingegebene Text mit einer für diese Lücke vorher festgelegten richtigen Antwort verglichen.

Informationen zum Erstellen einer Lückentext-Aufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 9.4 dargestellt.

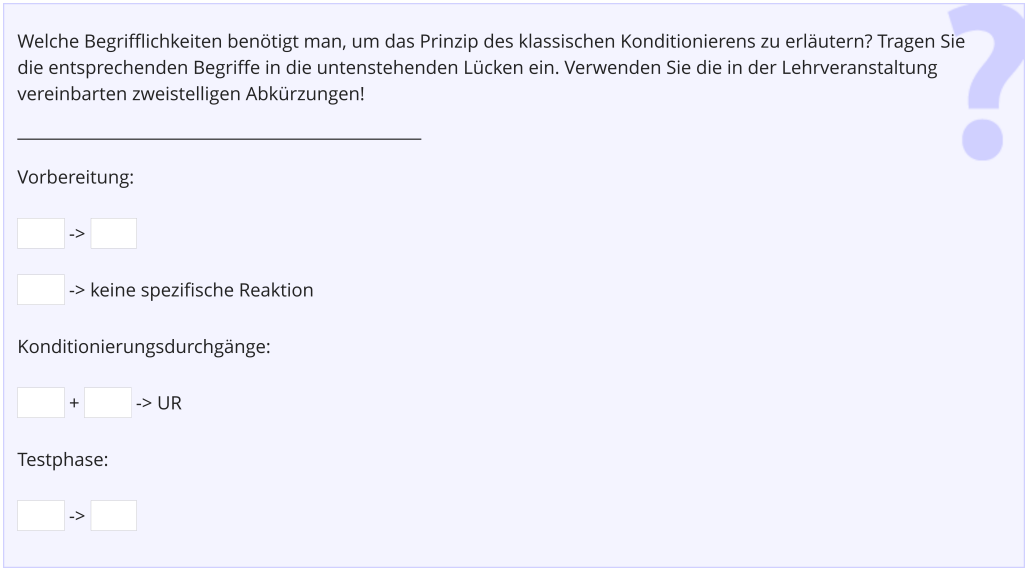


Abbildung 9.4. Beispiel einer Lückentext-Aufgabe vom Typ „Textlücke" in ILIAS. Die richtigen Antworten sind von oben nach unten: „US", „UR", „NS", „CS", „US", „CS", „CR".

9.2.5 *Long-Menu*-Frage

Long-Menu-Fragen sind eine Weiterentwicklung der Lückentext-Aufgaben. Der Unterschied besteht darin, dass beim Ergänzen der Textlücke den Prüflingen Antwortalternativen angeboten werden, sobald sie einige Buchstaben eingegeben haben. Ermöglicht wird das dadurch, dass vom Prüfenden zu jeder Textlücke eine – in der Regel „lange" – Liste von Antwortalternativen zur Verfügung gestellt wird. Für die automatische Bewertung ist bei jedem Element der Liste anzugeben, ob es eine korrekte Antwort oder ein Distraktor ist. Durch die Vorgabe von Alternativen sind die Antworten besser gegen Schreibfehler gesichert, die automatische Auswertung ist weniger fehleranfällig. Soll umgekehrt die korrekte Schreibweise mit geprüft werden, muss die Antwortliste auch möglichst alle möglichen Schreibfehler als Distraktoren enthalten. Alternativ zur automatischen Ergänzung beim Eingeben kann die Auswahlliste auch als langes *Drop-down*-Menü angeboten werden.

Informationen zum Erstellen einer *Long-Menu*-Frage stehen in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 9.5 dargestellt.

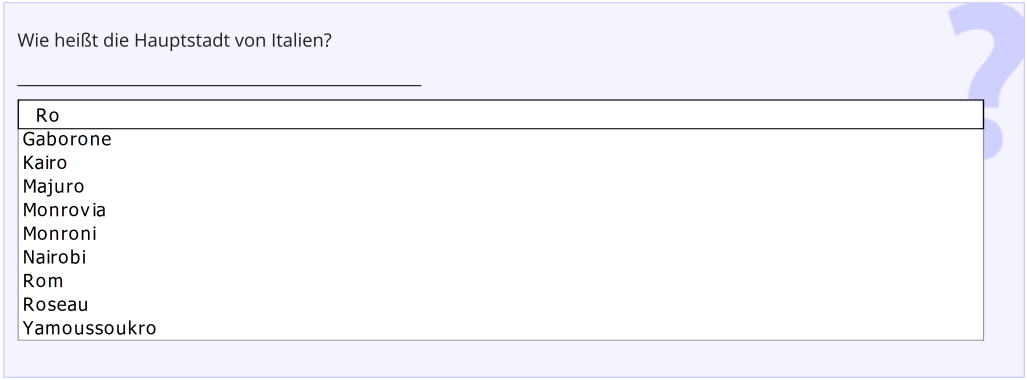


Abbildung 9.5. Beispiel einer *Long-Menu*-Frage in ILIAS. Die richtige Antwort ist hier aus einem langen Drop-down-Menü auszuwählen, wobei jede Eingabe im Antwortfeld als Filter auf die Listeneinträge wirkt. Für das Beispiel wurden über 200 Hauptstädte der Erde als Antwortalternativen hinterlegt. Durch die Eingabe der ersten beiden Buchstaben der richtigen Antwort „Rom" hat sich diese Liste bereits auf neun Einträge reduziert, in denen die Buchstabensequenz „Ro" ohne Beachtung der Groß- und Kleinschreibung vorkommt.

Ob man *Long-Menu*-Fragen als *single-response*- oder als offene Aufgaben auffasst, hängt offensichtlich von der Anzahl der Antwortalternativen in der Liste ab. Bei wenigen Optionen würde es sich eindeutig um *single-response*-Aufgaben handeln. Der Aufgabentyp *Long Menu* sollte daher nur dann in Betracht gezogen werden, wenn es sich tatsächlich um eine lange Liste von Antwortvorgaben handelt (mindestens 20, besser 50 oder mehr). Nur dann wird die Ratewahrscheinlichkeit so klein, dass man sie vernachlässigen und quasi von einer offenen Aufgabe sprechen kann. Andernfalls gelten für diesen Aufgabentyp die Ausführungen aus Kapitel 6.

9.2.6 Numerische Frage

Bei numerischen Fragen wird den Prüflingen eine Rechenaufgabe oder eine mit einer Zahl zu beantwortende Aufgabe präsentiert. Die numerische Antwort ist in das vorgegebene Antwortfeld einzutragen. Im Gegensatz zu Lückentext-Aufgaben vom Typ „numerische Lücke" (s. Abschnitt 9.2.3) kann bei numerischen Fragen nur eine einzige Zahl abgefragt werden.

ILIAS bietet bei diesem Aufgabentyp die Möglichkeit einer automatischen Bewertung. Dazu wird die durch die Prüflinge in der Lücke eingegebene Zahl mit einer für die Lücke vorher festgelegten richtigen Antwort verglichen. Es gibt die Möglichkeit, ein Intervall festzulegen, in dem sich die Lösung befinden darf.

Informationen zum Erstellen einer numerischen Frage stehen in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 9.6 dargestellt.

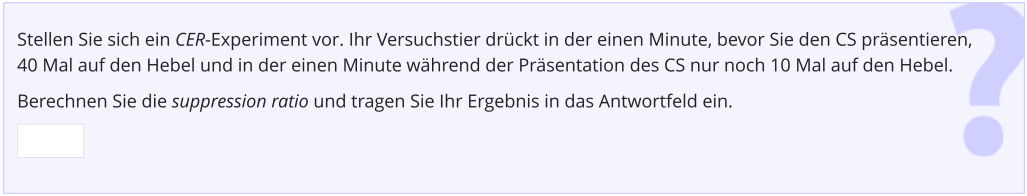


Abbildung 9.6. Beispiel einer numerischen Frage in ILIAS. Die richtige Antwort ist $SR = {}^{10}/_{(10+40)} = .20$.

9.2.7 Formelfrage

Bei einer Formelfrage wird den Prüflingen in der Regel eine Rechenaufgabe präsentiert, deren Struktur einer zuvor festgelegten Formel entspricht. Das Ergebnis der Berechnung ist als Zahl in ein vorgegebenes Antwortfeld einzutragen. Das besondere an einer Formelfrage ist, dass anstelle der in der Formel definierten Variablen automatisch Zufallszahlen generiert werden, die in einem vorher festgelegten Wertebereich liegen, so dass den Prüflingen immer wieder eine „neue“ Aufgabe präsentiert wird. Demzufolge geschieht die Auswertung auch automatisch durch ILIAS.

Informationen zum Erstellen einer Formelaufgabe stehen im Wiki des @LLZ zur Verfügung. Eine Dokumentation wird vom Entwickler bereitgestellt (Schottmüller, 2008). Ein einfaches Beispiel ist in Abbildung 9.7 dargestellt.

Abbildung 9.7. Beispiel einer Formelfragen-Aufgabe in ILIAS. Die richtigen Antworten sind „85“ und „0“.

9.2.8 *JSME*-Frage

Bei einer *JSME*-Frage haben die Prüflinge in der Regel die Aufgabe, die Struktur eines chemischen Moleküls abzubilden. Der Aufgabentyp wurde am @LLZ als Plugin für ILIAS entwickelt (Jobst & Annanias, 2014a) und verwendet den *Java Script Molecular Editor* (Bienfait & Ertl, 2013). Eine Dokumentation des Editors findet sich bei den Autoren. Ein einfaches Beispiel ist in Abbildung 9.8 dargestellt.

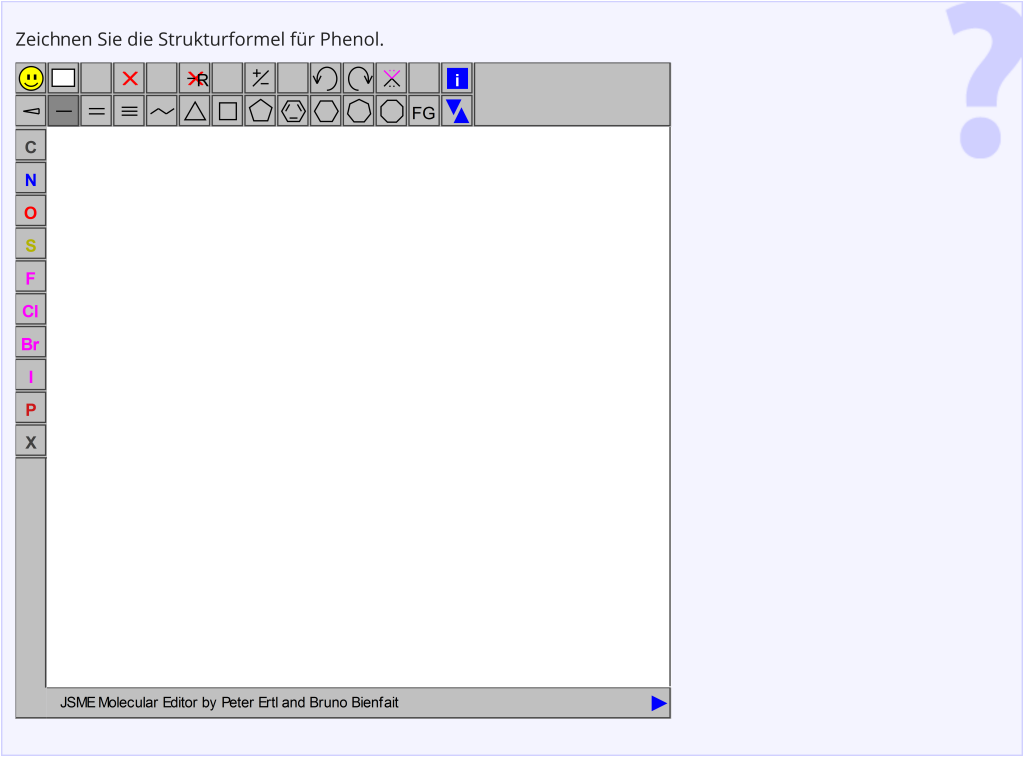


Abbildung 9.8. Beispiel einer *JSME*-Fragen-Aufgabe in ILIAS.

9.3 Parameter

Der einzige Parameter bei Aufgaben mit offenem Format ist die Anzahl der Punkte, die maximal erreicht werden können. Bei manchen Aufgabentypen (z. B. Lückentext oder Text-Teilmenge) wird die maximale Punktzahl im Regelfall übereinstimmen mit der Anzahl der Eingabefelder. Bei Freitext- oder JSME-Aufgaben wird die Anzahl der Punkte, die für eine vollständig richtige Lösung vergeben werden, von der Komplexität der Aufgabe abhängen. Eventuell ist es auch möglich, „Teilwissen" oder einzelne „Wissenselemente" zu identifizieren, für die jeweils Punkte vergeben werden, wenn sie aus der Aufgabenbeantwortung ersichtlich sind. In jedem Fall ist vom Prüfenden für jede Aufgabe festzulegen, wie viele Punkte maximal erreichbar sind und wie die Punkte vergeben werden.

Die Annahme einer Ratewahrscheinlichkeit hat bei Aufgaben mit offenem Format keinen rechten Sinn. Natürlich ist nicht auszuschließen, dass Prüflinge bei einer Aufgabe, die sie nicht fundiert beantworten können, irgendetwas antworten, eine spontane Assoziation oder einen Begriff, der im Kontext der Aufgabe in der Lehrveranstaltung häufiger vorkam oder etwas Ähnliches. Wenn damit in nennenswerter Weise korrekte Antworten produziert werden, ist das allerdings nicht ein Hinweis auf eine hohe Ratewahrscheinlichkeit, sondern eher darauf, dass in der Aufgabe nur dieses assoziative Wissen abgefragt wird. Reines Raten im Sinne der Antwortproduktion durch einen Zufallsgenerator führt bei freien Antwortformaten nur mit vernachlässigbar kleiner Wahrscheinlichkeit zu einer richtigen Antwort.

Vermutlich ist die Ratewahrscheinlichkeit beim freien Antwortformat sogar kleiner als der Fehler zweiter Art. Allerdings gilt auch hier: Wenn Prüflinge bei einer Aufgabe, die sie „eigentlich" gut beherrschen, Fehler machen, liegt das entweder am Prüfling, der unkonzentriert ist, nicht genau genug gelesen hat, die Aufgabe falsch verstanden hat etc. oder an der Aufgabenstellung, die missverständlich ist, nicht alle Möglichkeiten berücksichtigt, falsch formuliert ist etc. Im ersten Fall würde man wohl argumentieren, dass das Wissen doch nicht so sicher war und für die Beantwortung der Aufgabe nicht ausgereicht hat, im zweiten Fall ist es eher ein Problem der Aufgabenformulierung. Ob man dieses Problem mit einer festgelegten Toleranz für Flüchtigkeitsfehler würdigen soll, ist unseres Erachtens eine offene Frage. Es ist bislang bei Prüfungen nicht üblich, Flüchtigkeitsfehler explizit zu berücksichtigen und da wir bei offenen Aufgaben die Ratewahrscheinlichkeit mit $g = 0$ annehmen, erscheint es uns gerechtfertigt, hier auch für den Fehler zweiter Art eine Wahrscheinlichkeit von $f = 0$ anzunehmen.

9.4 Scoringverfahren

Bei Aufgabentypen mit einer automatischen Bewertung gibt es in der Regel keine Alternative zum sogenannten Standardverfahren: Für jede richtige Antwort, z. B. bei Text-Teilmengen-, Lückentext- oder Formel-Fragen, gibt es einen Punkt. Maluspunkte oder *testlet scoring* spielen beim freien Format wegen der verschwindenden Ratewahrscheinlichkeit keine Rolle. Bei den komplexeren Aufgabentypen wie Freitext- oder JSME-Fragen müssen die Punkte in der Regel manuell durch den Prüfenden vergeben werden, der je nach der Güte der Aufgabenlösung die volle Punktzahl oder Teilpunkte vergibt. Für die Transparenz des Scoringverfahrens ist es dabei hilfreich, wenn möglichst präzise definiert wird, welche Anforderungen an die Antwort gestellt werden und wofür es wie viele Punkte gibt.

9.5 Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte

Unter der Annahme $g = 0$ und $f = 0$ ist der Erwartungswert der in der Klausur erreichten Punkte nur vom Wissen p_W abhängig und es gilt die einfache Beziehung:

$$\mathcal{E}(X) = N \cdot p_W$$

mit:

$$N := \text{Anzahl der maximal erreichbaren Punkte.}$$

Die in der Klausur erreichten Punkte spiegeln also direkt das Wissen des Prüflings wider, wenn nur Aufgaben im freien Format verwendet werden. Dementsprechend müssen auch weder Bestehens- noch Notengrenzen korrigiert werden. In Abbildung 9.9 ist das beispielhaft für eine Klausur aus offenen Aufgaben mit insgesamt $N = 100$ erreichbaren Punkten wiedergegeben.

9.6 Zusammenfassung und Schlussfolgerung

Offene Aufgabenformate haben gegenüber Antwort-Wahl-Formaten viele Vorteile. Zum einen hat man bei der Erstellung der Aufgaben einen geringeren Vorbereitungsaufwand. Man

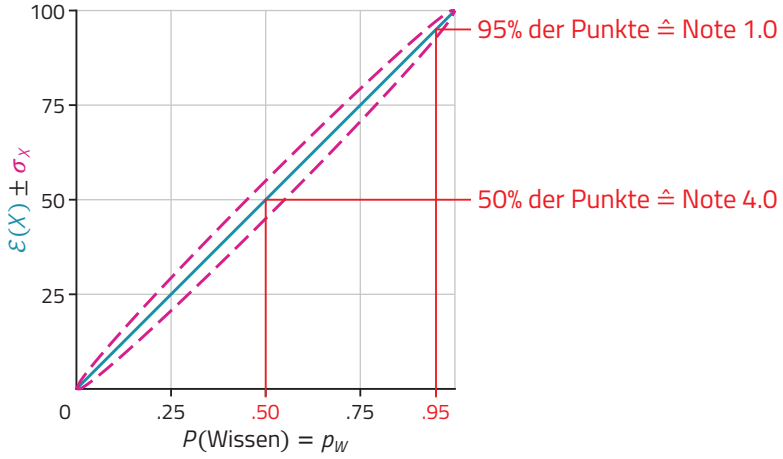


Abbildung 9.9. Bestehens- und Notengrenzen in Abhängigkeit vom Wissen beim Standardscoring. Beispiel für eine Prüfung aus offenen Aufgaben mit insgesamt $N = 100$ erreichbaren Punkten ($g = 0$, h ist unbestimmt, $f = 0$). Die Punktwerte der Bestehensgrenze (Note 4.0) bei einem Wissen von $p_W = .50$ und der Bestnotengrenze (Note 1.0) bei einem Wissen von $p_W = .95$ sind rechts in rot angegeben.

muss sich beispielsweise keine gut passenden, aber eindeutig falschen Distraktoren ausdenken. Dies kann insbesondere bei *single-response*-Aufgaben einen erheblichen zeitlichen und organisatorischen Aufwand darstellen und erfordert große Sorgfalt, wenn die Aufgabe das zu prüfende Wissen adäquat widerspiegeln soll. Zum anderen ist man bei der Formulierung offener Aufgaben kaum eingeschränkt, da sich die in ILIAS integrierten Aufgabentypen sehr flexibel und kreativ einsetzen lassen.

Ein weiterer Vorteil besteht darin, dass die Ratewahrscheinlichkeit offener Aufgabenformate in aller Regel gegen null geht. Es ist bei diesem Aufgabenformat so gut wie unmöglich, bei Nichtwissen durch blindes Raten die richtigen Antworten zu produzieren, sofern der Suchraum nicht zu weit eingeschränkt ist, z. B. durch eine zu enge Formulierung der Aufgabe. Die geringe Ratewahrscheinlichkeit äußert sich bei der Bewertung der Aufgaben darin, dass die zu erwartenden Punkte direkt das Wissen der Prüflinge widerspiegeln – zumindest dann, wenn man auch die Flüchtigkeitsfehlertoleranz mit $f = 0$ annimmt.

Demgegenüber haben offene Aufgaben im Allgemeinen eine geringere Auswertungsobjektivität als Aufgaben im Antwort-Wahl-Verfahren. Die Bewertung offener Aufgaben, bei denen die Prüflinge nicht aus vorgegebenen Antwortalternativen auswählen, sondern ihre Antworten frei formulieren, ist sehr viel komplexer und in der Regel von der subjektiven Einschätzung

des Prüfenden abhängig. Weiterhin muss die Auswertung von offenen Aufgaben, auch größtenteils in ILIAS, manuell erfolgen und nimmt dadurch deutlich mehr Zeit in Anspruch als die automatische Auswertung von Aufgaben im Antwort-Wahl-Verfahren.

Fragen im offenen und im geschlossenen Antwortformat können in einer Klausur beliebig gemischt werden. Das Vorgehen bei der Auswertung von Klausuren mit unterschiedlichen Aufgabenformaten und insbesondere die Berechnung der ratekorrigierten Bestehens- und Notengrenzen für diesen Fall wird in Kapitel 12 behandelt und genauer dargestellt.

10 Aufgaben mit abhängigen Antwortalternativen

10.1 Charakteristik

Zu den Aufgaben mit abhängigen Antwortalternativen zählen die Zuordnungs- und die Anordnungsaufgaben. Bei Zuordnungsaufgaben werden den Prüflingen zwei Listen mit Begriffen vorgegeben. Jedem Begriff der ersten Liste (in ILIAS nennt man sie „Definitionen") ist dabei genau ein Begriff der zweiten Liste (in ILIAS: „Terme") zuzuordnen, wobei Terme nicht mehrfach zugeordnet werden dürfen. Bei Anordnungsaufgaben ist eine beliebige Anzahl an Begriffen vorgegeben, die von den Prüflingen in die richtige Reihenfolge gebracht werden sollen.

Die Abhängigkeit zwischen den Zuordnungen kommt dadurch zustande, dass jeder Term nur höchstens einmal verwendet werden darf. Wird z. B. der ersten Definition fälschlicherweise der Term zugeordnet, der eigentlich zur zweiten Definition gehört, dann gibt es bei dieser zweiten Definition keine richtige Antwort mehr. Diese Abhängigkeit ist besonders problematisch, wenn Prüflinge aufgrund fehlenden Wissens raten. Raten sie richtig, erhöht sich die Wahrscheinlichkeit, weitere Zuordnungen richtig zu raten. Raten sie falsch, besteht die Möglichkeit, dass damit automatisch ein weiterer Fehler erzwungen wird.

Die Abhängigkeit wäre aufgelöst, wenn zugelassen würde, dass ein Term mehreren Definitionen zugeordnet wird. Für jede Definition könnte dann der zugehörige Begriff aus der gesamten Liste der Terme ausgewählt werden. Die Aufgabe wäre in diesem Fall allerdings besser als eine Folge von *single-response*-Aufgaben aufzufassen, die in Kapitel 6 beschrieben sind. Als Zuordnungsaufgaben bezeichnen wir deshalb hier nur Aufgaben, bei denen jeder Term nur höchstens einmal verwendet, also höchstens einem Begriff zugeordnet werden darf. Dabei kann die Anzahl der zur Verfügung stehenden Terme durchaus höher sein als die Anzahl der Definitionen. Ist die Anzahl der Terme gleich der Anzahl der Definitionen, so dass alle Terme zugeordnet werden, spricht man von einer 1-zu-1-Zuordnung.

Anordnungsaufgaben sind formal sehr ähnlich zu Zuordnungsaufgaben. Die vorgegebenen Begriffe werden hier Rangplätzen zugeordnet. Allerdings sind die kognitiven Anforderungen dieser beiden Aufgabentypen sehr unterschiedlich. Während es bei den Zuordnungsaufgaben um die Beziehung zwischen Definitionen und Termen geht, wird bei Anordnungsfragen nach der Ordnungsrelation der Begriffe untereinander gefragt. Die beiden Aufgabentypen müssen deshalb trotz ihrer formalen Ähnlichkeit getrennt betrachtet werden und haben sehr unterschiedliche Eigenschaften.

Gemeinsam ist beiden Aufgabentypen, dass sie mehrere Fragen in einer Aufgabe zusammenfassen und damit eine sehr kompakte Aufgabenstellung ermöglichen: Man kann z. B. gleichzeitig die Hauptstädte mehrerer Länder abfragen, Bauteile einer Maschine benennen lassen, die zeitliche Abfolge von historischen Ereignissen erfragen oder wässrige Lösungen nach ihrem pH-Wert anordnen lassen. Der Preis für die Ökonomie bei dieser Aufgabenstellung ist allerdings, dass oft nicht ganz eindeutig ist, welches Wissen für eine richtige Antwort nötig ist. Die Zusammenfassung mehrerer Fragen zu einer Aufgabe führt zu den bereits erwähnten Abhängigkeiten beim Beantworten und damit zu komplizierteren Auswertungsschemata. Die Behandlung der Anordnungs- und Zuordnungsaufgaben in den folgenden Abschnitten ist deshalb an manchen Stellen noch etwas technischer als in den früheren Abschnitten. Der Aufwand für eine genaue Analyse ist aber lohnend. Sie zeigt nämlich, dass dieser Aufgabentyp komplexer ist, als es auf den ersten Blick erscheinen mag, dass die Ratewahrscheinlichkeit höher ist, als man bei naiver Betrachtung vermuten würde und deshalb eine (angemessene) Ratekorrektur unerlässlich ist.

10.2 Aufgaben mit abhängigen Antwortalternativen in ILIAS

In ILIAS sind den Aufgaben mit abhängigen Antwortalternativen folgende Aufgabentypen zuzuordnen:

- Zuordnungsfrage
- Anordnungsfrage
- Anordnungsfrage (horizontal)

10.2.1 Zuordnungsfrage

Bei einer Zuordnungsaufgabe wird den Prüflingen auf der linken Seite eine Reihe von Definitionen oder Bildern präsentiert, denen mittels *Drag & Drop* Terme von der rechten Seite zugeordnet werden müssen. Wählt man den Zuordnungsmodus „1:1“, müssen die Zuordnungen eindeutig sein, d. h., dass jedem Bild oder jeder Definition genau ein Term zugeordnet werden muss und kein Term mehrfach verwendet werden kann. Es besteht die Möglichkeit, überzählige Terme anzubieten, die zu keiner Definition passen und als Distraktoren dienen.

Darüber hinaus gibt es seit ILIAS 5.0 den Zuordnungsmodus „n:n“ bei dem jeder Term mehrfach und jeder Definition mehr als ein Term zugeordnet werden kann. Dieser Modus ist sehr flexibel und bietet die Möglichkeit, auch andere Aufgabenformate (s. Kapitel 5) darzustellen. So könnte man z. B. eine *multiple-true-false*-Aufgabe als Zuordnungsaufgabe erstellen, indem die Terme einer Liste zwei Kategorien zugeordnet werden, die durch die Definitionen gegeben sind. Die jeweiligen Entscheidungen, die von den Prüflingen zu treffen sind, müssen dann nicht mehr voneinander abhängig sein. Da dieses Kapitel Aufgaben mit abhängigen Antwortalternativen betrachtet, beziehen sich die Überlegungen und Berechnungen auf den „1:1“-Zuordnungsmodus.

Informationen zum Erstellen einer Zuordnungsaufgabe stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 10.1 dargestellt.

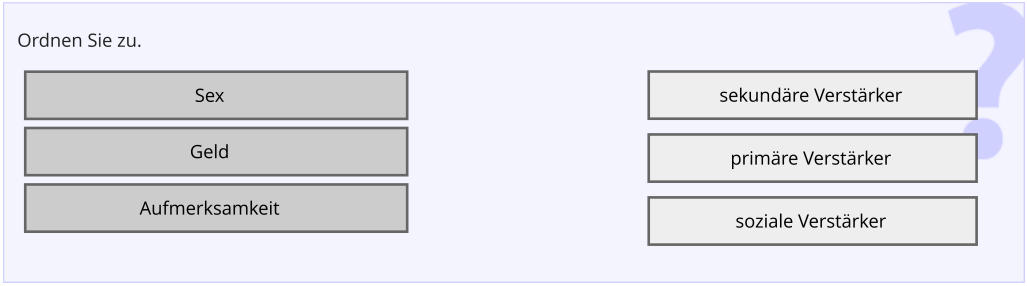


Abbildung 10.1. Beispiel einer Zuordnungsfrage in ILIAS mit drei Termen und drei Definitionen. Die richtige Lösung ist, die Terme so zu verschieben, dass sich folgende Zuordnungen zu den Begriffen ergeben: „Aufmerksamkeit“ und „soziale Verstärker“, „Geld“ und „sekundäre Verstärker“ sowie „Sex“ und „primäre Verstärker“.

10.2.2 Anordnungsfrage

Bei einer Anordnungsaufgabe wird den Prüflingen eine Reihe von Bildern oder Begriffen präsentiert. Je nach Aufgabenstellung müssen sie diese mittels *Drag & Drop* aufsteigend oder absteigend sortieren.

Informationen zum Erstellen einer Anordnungsfrage stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 10.2 dargestellt.

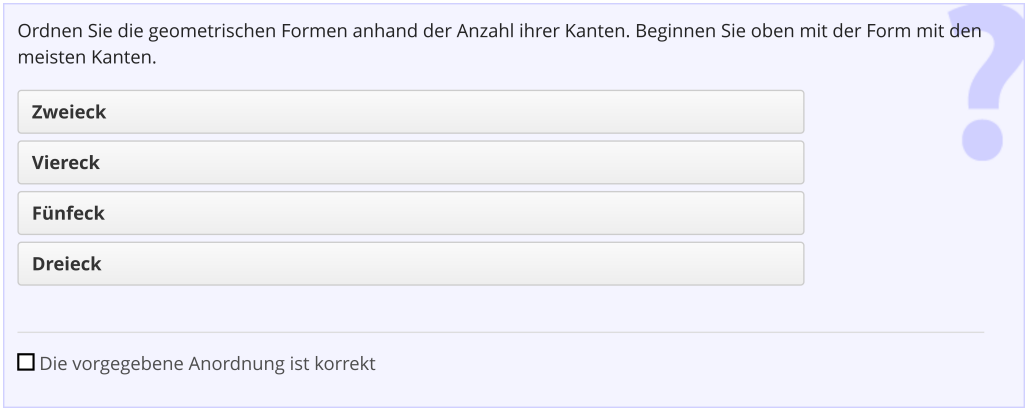


Abbildung 10.2. Beispiel einer Anordnungsfrage in ILIAS mit vier Begriffen. Die richtige Lösung ist, die Begriffe so zu verschieben und anzuordnen, dass sich von oben nach unten die Reihenfolge „Fünfeck", „Viereck", „Dreieck" und „Zweieck" ergibt.

10.2.3 Anordnungsfrage (horizontal)

Bei einer horizontalen Anordnungsaufgabe wird den Prüflingen eine Reihe von Begriffen präsentiert. Die Aufgabe der Prüflinge ist es, diese Begriffe mittels *Drag & Drop* je nach Aufgabenstellung von links nach rechts in die richtige Reihenfolge zu bringen.

Informationen zum Erstellen einer horizontalen Anordnungsfrage stehen im Wiki des @LLZ und in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 10.3 dargestellt.

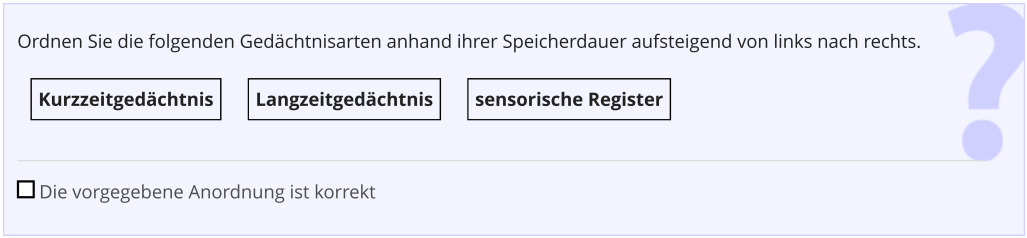


Abbildung 10.3. Beispiel einer horizontalen Anordnungsfrage in ILIAS mit drei Begriffen. Die richtige Lösung ist, die Begriffe so zu verschieben und anzuordnen, dass sich von links nach rechts die Reihenfolge „sensorische Register", „Kurzzeitgedächtnis" und „Langzeitgedächtnis" ergibt.

10.3 Parameter

10.3.1 Anzahl der Definitionen *k*

Die Anzahl der Definitionen, d. h. der Begriffe, denen bei Zuordnungsaufgaben ein Term zuzuordnen ist, wird mit k bezeichnet. Der Wert von k bestimmt hier auch die Anzahl der Entscheidungen, die zu treffen sind. Bei Anordnungsaufgaben bezeichnet k die Anzahl der anzuordnenden Begriffe.

10.3.2 Anzahl der Terme *m* (nur bei Zuordnungsfragen)

Die Anzahl der den Definitionen zuzuordnenden Terme bei Zuordnungsfragen wird mit m bezeichnet. Der Wert von m bestimmt damit die Anzahl der Antwortmöglichkeiten zu Beginn des Antwortprozesses, also bei der Auswahl der ersten Zuordnung. Allgemein gilt $m \geq k$, bei 1-zu-1-Zuordnungen ist $m = k$. Die $m - k$ „überzähligen" Terme werden als Distraktoren bezeichnet.

10.3.3 Ratewahrscheinlichkeit *g*

Bei Zu- und Anordnungsaufgaben sind typischerweise mehrere Entscheidungen zu treffen, die jeweils falsch oder richtig sein können. Die Ratewahrscheinlichkeit für die Beantwortung der Aufgabenteile ist dabei nicht konstant, sondern verändert sich nach jedem einzelnen Schritt und ist insbesondere abhängig von den bereits erfolgten (Teil-)Antworten. Da der

Antwortprozess für die beiden hier betrachteten Aufgabenformate sehr unterschiedlich ist, müssen sie getrennt behandelt werden.

Zuordnungsaufgaben

Eine Zuordnungsaufgabe mit k-vielen Definitionen und m-vielen Termen bezeichnen wir als (m, k)-Aufgabe mit $m \geq k$. Jeder Definition kann entweder der richtige oder ein falscher Term zugeordnet werden. Bei jeder (m, k)-Aufgabe gibt es insgesamt $m!/(m-k)!$ unterschiedliche Zuordnungsmuster. Bei wie vielen davon 0, 1, 2, … oder alle k Zuordnungen richtig sind, lässt sich über die Theorie der fixpunktfreien Permutationen angeben (z. B. Diekert, Kufleitner & Rosenberger, 2013). Die Anzahl aller möglichen Zuordnungsmuster mit genau y-vielen richtigen Zuordnungen lässt sich allgemein bestimmen durch:

$$D(m, k, y) := \frac{\binom{k}{y}}{(m-k)!} \cdot \sum_{i=0}^{k-y} (-1)^i \cdot \binom{k-y}{i} \cdot (m-y-i)! \,. \tag{10.1}$$

Eine kombinatorische Herleitung dieser Formel findet man z. B. bei Hanson, Seyffarth und Weston (1983). Die Anwendung auf wahrscheinlichkeitstheoretische Überlegungen bei Zuordnungsaufgaben diskutieren Jančařík und Kostelecká (2015).

Für den Spezialfall einer 1-zu-1-Zuordnung mit $m = k$ vereinfacht sich Gleichung 10.1 zur Formel für die sogenannten Rencontres-Zahlen (Jacobs & Jungnickel, 2004):

$$D(k, y) = \frac{k!}{y!} \cdot \sum_{i=0}^{k-y} \frac{(-1)^i}{i!}. \tag{10.2}$$

Aus Gleichung 10.1 lassen sich auch die Wahrscheinlichkeiten für $y = 0, 1, \ldots, k$-viele richtige Antworten bei „blindem Raten", also bei einer zufälligen (im Sinne von: gleichverteilten) Auswahl von Zuordnungen berechnen, indem man $D(m, k, y)$ jeweils durch die Anzahl aller möglichen Antwortmuster dividiert:

$$P(Y = y \mid W = 0) = \frac{(m-k)!}{m!} \cdot D(m, k, y).$$

Für eine (m, k)-Aufgabe bezeichnet dabei die Zufallsvariable Y die Anzahl der richtigen Zuordnungen und die Zufallsvariable W bezeichnet die Anzahl der Zuordnungen, die gewusst werden. $W = 0$ heißt also: Der Studierende kennt keine einzige Antwort und muss alle k

Tabelle 10.1. Bedingte Wahrscheinlichkeitsverteilung der Anzahl der richtigen Zuordnungen in Abhängigkeit vom Wissen. Angegeben ist die Wahrscheinlichkeit $P(Y = y \mid W = w)$, d. h. y-viele Zuordnungen korrekt zu lösen, wenn w-viele Zuordnungen gewusst und die übrigen $k - w$ Zuordnungen geraten werden. Beispiel für eine 1-zu-1-Zuordnungsaufgabe mit $k = 4$ zuzuordnenden Begriffspaaren.

y	0	1	2	3	4
$w = 0$	.375	.333	.250	0*	.042
$w = 1$	0	.333	.500	0*	.167
$w = 2$	0	0	.500	0*	.500
$w = 3$	0	0	0	0*	1
$w = 4$	0	0	0	0*	1

* Bei $k = 4$ zuzuordnenden Begriffspaaren ist es unmöglich, nur drei Zuordnungen korrekt zu lösen, da bei drei richtigen als letzte Zuordnung nur die richtige übrig bleibt.

Zuordnungen raten.

Der allgemeinere Fall, dass nämlich w-viele Zuordnungen bekannt sind ($w = 0, 1, \ldots, k$), lässt sich analog behandeln. Wenn wir von Flüchtigkeitsfehlern absehen (s. Abschnitt 10.3.4) nehmen wir an, dass zunächst die w-vielen gewussten Zuordnungen richtig beantwortet werden, und dann bei den restlichen $(k - w)$-vielen Zordnungen geraten wird. Die Wahrscheinlichkeitsfunktion für die Anzahl richtiger Antworten Y bei einer (m, k)-Zuordnungsfrage lautet dann für alle $y = 0, 1, \ldots k$ und $w = 0, 1, \ldots k$:

$$P(Y = y \mid W = w) = \begin{cases} 0 & \text{falls } y < w \\ \frac{(m-k)!}{(m-w)!} \cdot D(m - w, k - w, y - w) & \text{sonst.} \end{cases} \tag{10.3}$$

Beispielhaft sind in Tabelle 10.1 für eine 1-zu-1-Zuordnungsaufgabe mit $k = 4$ Begriffspaaren die Wahrscheinlichkeiten dafür angegeben, y-viele Zuordnungen korrekt zu lösen, wenn man w-viele Zuordnungen weiß. Aus dieser Tabelle ist auch ersichtlich, dass es für die vollständige, richtige Beantwortung ausreicht, drei der vier Zuordnungen zu kennen. Wenn man die Hälfte der Zuordnungen kennt ($w = 2$), beträgt die Wahrscheinlichkeit für eine komplett richtige Antwort bereits 1/2.

Anordnungsaufgaben

Bei Anordnungsaufgaben kann man sich zwar eine ähnliche Vorstellung vom Lösungsprozess wie bei Zuordnungsaufgaben machen – die Prüflinge ordnen erst diejenigen Elemente an, deren

Relation zueinander sie kennen und sortieren dann die restlichen Elemente zufällig irgendwo ein –, allerdings sind die einzelnen Schritte in diesem Prozess nicht eindeutig identifizierbar und als richtig oder falsch bewertbar (s. Abschnitt 10.4.2). In der Regel wird deshalb eine Anordnungsaufgabe wie eine Einzelaufgabe behandelt, die genau dann richtig beantwortet ist, wenn die Reihenfolge aller Elemente korrekt angegeben wurde. Bei einer zufälligen Anordnung von k-vielen Elementen ohne jedes Wissen im Sinne der Gleichverteilung aller Möglichkeiten beträgt die Ratewahrscheinlichkeit $g = 1/k!$.

10.3.4 Flüchtigkeitsfehler *f*

Auch bei Anordnungs- und Zuordnungsaufgaben muss prinzipiell damit gerechnet werden, dass Prüflinge trotz sicheren Wissens an der einen oder anderen Stelle eine falsche Antwort geben. Die Auswirkungen einer falschen Entscheidung sind wegen der in Abschnitt 10.1 beschriebenen Abhängigkeiten sogar besonders gravierend, weil sie weitere Fehler zur Folge haben kann. Allerdings ist gerade durch diese Abhängigkeit auch die Chance groß, einen Flüchtigkeitsfehler rechtzeitig zu entdecken und zu korrigieren, z. B. dadurch, dass für eine Definition kein passender Term gefunden wird und deshalb die bisherigen An- bzw. Zuordnungen überprüft werden.

Wir gehen bei Flüchtigkeitsfehlern davon aus, dass Prüflinge tatsächlich über (Teil-)Wissen verfügen und nur aus Unachtsamkeit eine falsche An- bzw. Zuordnung treffen. Die explizite Berücksichtigung von Flüchtigkeitsfehlern ist deshalb nicht ohne Weiteres möglich und würde weitere – zum Teil weitgehende – Annahmen über den Prozess des Lösungsverhaltens erfordern. Die Wahrscheinlichkeit für einen Flüchtigkeitsfehler wird deshalb mit dem konstanten Wert 0 angenommen ($f = 0$), so dass in den formalen Modellierungen für die Anordnungs- und Zuordnungsaufgaben kein Term für einen Flüchtigkeitsfehler enthalten ist. Aus den gerade genannten Gründen ist dies unseres Erachtens vertretbar.

10.3.5 Rateneigung *h*

Für ähnlich unkritisch halten wir die Festlegung der Rateneigung h auf den Wert $h = 1$ bei den Zuordnungs- und Anordnungsaufgaben. Abgesehen davon, dass ohnehin in der Regel die Prüflinge aufgefordert werden, *alle* Aufgaben zu bearbeiten und im Zweifel zu raten (vgl. dazu die Empfehlungen in Kapitel 4), gibt es bei Anordnungsaufgaben in der gängigen elektronischen

Darstellung gar nicht die Möglichkeit, einzelne Begriffe nicht einzuordnen. Unternehmen Prüflinge nichts, wird die ursprünglich vorgegebene Position der Begriffe als Antwort übernommen. Das entspricht bei zufälliger Anordnung der Begriffe blindem Raten.

Bei Zuordnungsaufgaben ist ein Auslassen von Antworten technisch zwar möglich, würde aber erst Sinn ergeben, wenn – bei 1-zu-1-Zuordnungen – mindestens zwei Zuordnungen ausgelassen werden. Und: keine Antwort oder eine unvollständige Antwort zu geben ist bei allen in Frage kommenden Scoringverfahren (s. Abschnitt 10.4) eine schlechtere Strategie als Raten. Die Rateneigung wird daher in den Berechnungen dieses Kapitels konstant auf den Wert $h = 1$ gesetzt und kommt in den Formeln nicht mehr explizit als Parameter vor.

10.4 Scoringverfahren

10.4.1 Zuordnungsaufgaben

Bei den Zuordnungsaufgaben kann für jede der k-vielen Zuordnungen entschieden werden, ob sie richtig oder falsch ist. Prinzipiell kommt deshalb sowohl das Standardverfahren (ein Punkt für jede richtige Zuordnung) als auch ein *testlet-scoring*-Verfahren (Punkte für die Gesamtlösung einer Aufgabe) in Frage. Die Auswirkungen beider Verfahren werden in Abschnitt 10.5.1 genauer analysiert. Maluspunktverfahren mit negativen Punktwerten für falsche Antworten sind dagegen bei Zuordnungsaufgaben nicht sinnvoll. Der wichtigste Grund dafür ist, dass die Ratewahrscheinlichkeiten nicht konstant sind und damit die Grundlage für das *formula scoring* fehlt, alle anderen Maluspunktverfahren aber nicht gut begründbar sind (s. Abschnitt 3.2).

10.4.2 Anordnungsaufgaben

Bei den Anordnungsaufgaben ist nur eine Gesamtbewertung der Anordnung als insgesamt richtig oder falsch sinnvoll. Für die Vergabe von Punkten für Teillösungen gibt es kein unmittelbar überzeugendes Verfahren. Man könnte zwar z. B. auszählen, wie viele Begriffe an der richtigen Stelle, also auf dem richtigen Rangplatz, stehen, das ist aber offensichtlich kein zuverlässiges Maß für die Güte der (Teil-)Lösung. Setzt man z. B. den letzten Begriff fälschlicherweise an die erste Stelle und ordnet alle anderen Begriffe korrekt, dann ist vieles an dieser Lösung richtig, es stehen aber alle Begriffe auf dem falschen Rangplatz.

Eine plausiblere Möglichkeit für die Honorierung von Teillösungen wäre es, die Anzahl der Inversionen zu bestimmen, also die Anzahl der Begriffspaare, die in der falschen Reihenfolge angeordnet wurden, oder die minimale Anzahl der Vertauschungen benachbarter Begriffe, die nötig sind, um die korrekte Ordnung herzustellen. Ob damit eine sinnvolle Bewertung der Aufgabenlösung im Sinne der Aufgabenstellung gelingt, ist aber eher fraglich. In ILIAS (Stand: Version 5.1) ist dementsprechend bei Anordnungsaufgaben auch nur eine Gesamtbewertung als richtig oder falsch möglich.

10.5 Zusammenhang zwischen dem Wissen p_W und dem Erwartungswert der Punkte

Der Zusammenhang zwischen dem Wissen p_W eines Prüflings und dem Erwartungswert der Zufallsvariablen X („erreichte Punkte in der Klausur") wird uns wieder wichtige Kriterien dafür liefern, inwieweit Anordnungs- und Zuordnungsaufgaben für Klausuren geeignet sind, welche Auswirkungen die Ratewahrscheinlichkeit auf das Klausurergebnis hat, wie sich unterschiedliche Scoringverfahren auswirken und welche Ratekorrekturen bei den Bestehens- und Notengrenzen gegebenenfalls möglich sind. Anordnungs- und Zuordnungsaufgaben müssen dabei wegen ihrer unterschiedlichen Eigenschaften wieder getrennt betrachtet werden.

10.5.1 Zuordnungsaufgaben

Wir gehen wieder davon aus, dass eine Klausur mit n-vielen (m, k)-Zuordnungsaufgaben vorliegt. Bei jeder Aufgabe muss jedem der k Begriffe genau einer der m Terme zugeordnet werden. Mit Y_i bezeichnen wir die Zufallsvariable „erreichte Punkte bei Aufgabe i".

Standardscoring

Beim Standardscoring gibt es für jede richtige Zuordnung einen Punkt, bei jeder (m, k)-Aufgabe sind also maximal k-viele Punkte möglich. Die Maximalpunktzahl in der Klausur beträgt damit $N := n \cdot k$. Für den Erwartungswert von X gilt:

$$\mathcal{E}(X) = \sum_{i=1}^{n} \mathcal{E}(Y_i) = n \cdot \mathcal{E}(Y),$$

so dass wir uns auf den Erwartungswert $\mathcal{E}(Y)$ bei einer (m, k)-Aufgabe beschränken können:

$$\begin{aligned}\mathcal{E}(Y) &= \sum_{y=0}^{k} y \cdot P(Y = y) \\ &= \sum_{y=0}^{k} \sum_{w=0}^{k} y \cdot P(Y = y, W = w) \\ &= \sum_{y=0}^{k} \sum_{w=0}^{k} y \cdot P(Y = y \mid W = w) \cdot P(W = w). \end{aligned} \tag{10.4}$$

Für den Fall $p_W = 0$, wenn also nur geraten wird, nimmt dieser Erwartungswert eine besonders einfache Form an. Für jede (m, k)-Aufgabe gilt in diesem Fall nach Hanson et al. (1983, S. 229):

$$\mathcal{E}(Y) = \frac{k}{m}. \tag{10.5}$$

Für jede 1-zu-1-Zuordnungsaufgabe mit k-vielen Termen und Definitionen nimmt der Erwartungswert $\mathcal{E}(Y)$ immer den Wert 1 an und ist damit etwas überraschend unabhängig von k. Wer bei Zuordnungsaufgaben blind rät, hat im Schnitt eine richtige Zuordnung bei Aufgaben mit $k = m$. Wenn es zusätzlich Distraktor-Terme gibt ($m > k$), dann liegt der Erwartungswert sogar darunter.

Im allgemeinen Fall beliebiger Werte für das Wissen p_W lässt sich der Erwartungswert von Y über eine Auswertung von Gleichung 10.4 bestimmen, da die beiden Wahrscheinlichkeitsfunktionen auf der rechten Seite der Gleichung bekannt sind: $P(Y = y \mid W = w)$ ist durch Gleichung 10.3 definiert und die Zufallsvariable W, die Anzahl der Zuordnungen, die ein Prüfling bei einer (m, k)-Aufgabe weiß, ist binomialverteilt mit den Parametern k und p_W. Dabei ist p_W die Wahrscheinlichkeit dafür, dass ein Prüfling die richtige Antwort weiß (im Sinne des Wahrscheinlichkeitsmodells aus Abschnitt 2.2), bezogen auf jede einzelne Zuordnung.

Zur Erinnerung: wir betrachten p_W als einen Wert, der das Wissen eines Prüflings charakterisiert und daher für jeden Prüfling als Konstante aufgefasst wird. Das zugrundeliegende Zufallsexperiment besteht darin, dass Prüfende zufällig eine Aufgabe aus der Menge aller zum Prüfungsstoff gehörenden Aufgaben ziehen und ein Prüfling die Antwort darauf weiß oder nicht weiß. Die einzelnen Aufgaben können dabei unabhängig voneinander gewusst oder nicht gewusst werden, wenn Prüfende nicht inhaltliche Abhängigkeiten durch ungeschickte Formulierungen erzeugen, so dass bei 1-zu-1-Zuordnungsaufgaben durchaus auch drei von

vier Zuordnungen gewusst werden können. Die bei dieser Aufgabenart typische Abhängigkeit entsteht erst beim Beantworten der Aufgaben: wer drei von vier Aufgaben weiß, beantwortet auch die vierte richtig, weil nach drei richtigen Antworten die Ratewahrscheinlichkeit g den Wert 1 annimmt. Es gilt also für alle $w = 0, 1, \ldots, k$:

$$P(W = w) = \binom{k}{w} \cdot {p_W}^w \cdot (1 - p_W)^{k-w}. \tag{10.6}$$

Das Einsetzen der Gleichungen 10.3 und 10.6 in Gleichung 10.4 ermöglicht die Berechnung des Erwartungswertes von Y und damit auch von X als Funktion von p_W. In Abbildung 10.4 sind links die Erwartungswertfunktionen für verschiedene Werte von k in 1-zu-1-Zuordnungsaufgaben dargestellt und rechts die entsprechenden Funktionen für verschiedene Werte von m bei $(m, 4)$-Aufgaben. Interessant sind dabei vor allem die folgenden Beobachtungen:

- Die Erwartungswertkurven beider Abbildungen starten am linken Rand ($p_W = 0$) jeweils mit dem Wert $^{100}/_{m}$. D. h., wer ohne jedes Wissen blind rät, kann bei einer Klausur, die nur (m, k)-Aufgaben enthält, in Übereinstimmung mit Gleichung 10.5 im Schnitt $(^{1}/_{m}) \cdot 100\%$ der maximalen Punkte erreichen.
- Alle Kurven enden am rechten Rand ($p_W = 1$) bei einem Wert von 100. Da wir den Fehler zweiter Art mit $f = 0$ angenommen haben, erhalten Prüflinge, die alles wissen, auch die maximale Punktzahl.
- Zwischen diesen beiden Extrempunkten steigen die Kurven mit wachsendem Wissen monoton an, allerdings nicht linear, sondern mit abnehmender Steigung.

Testlet Scoring

Beim *testlet scoring* werden Punkte nicht für jede einzelne Zuordnung vergeben, sondern für die Güte der Gesamtlösung. Wir betrachten hier nur den Fall, der in Abschnitt 3.3.1 als „Alles-oder-nichts"-Verfahren (AoN-Verfahren) bezeichnet wurde. Dabei gibt es Punkte nur für vollständig richtige Zuordnungen. Denkbar wären hier als Alternative auch Zwischenlösungen, z. B. im Sinne eines K'-Verfahrens mit halben Punkten für „fast" richtige Lösungen oder ähnlicher Regeln für die Punktvergabe. Die Konsequenzen solch spezieller Verfahren können im Bedarfsfall mit den hier angewandten Methoden leicht untersucht werden. Dabei dürften sich aber kaum überraschende Ergebnisse zeigen. Wir beschränken uns deshalb hier auf die Analyse des AoN-Verfahrens und nehmen an, dass bei einer (m, k)-Aufgabe k-viele Punkte

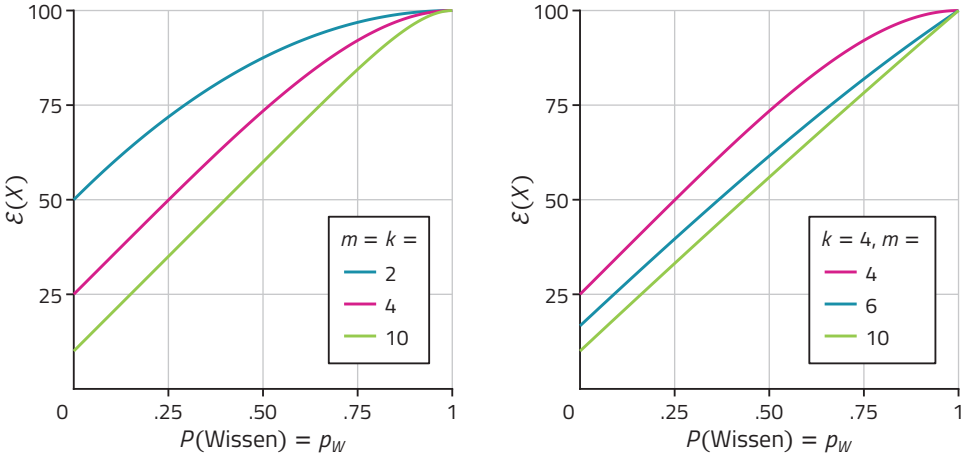


Abbildung 10.4. Erwartungswert für eine Klausur mit n-vielen Aufgaben gleichen Typs als Funktion vom p_W beim Standard-Scoring. Die Anzahl der Aufgaben n wurde jeweils so gewählt, dass immer genau $N = 100$ Punkte maximal in der Klausur zu erreichen waren. Man kann die Ordinate deshalb auch interpretieren als: „Erwartungswert der Punkte in Prozent der Maximalpunktzahl". Links: Erwartungswertfunktionen für eine Klausur mit (k, k)-Zuordnungsaufgaben (von oben nach unten: $k = 2, 4, 10$). Rechts: Erwartungswertfunktionen für eine Klausur mit $(m, 4)$-Zuordnungsaufgaben (von oben nach unten: $m = 4, 6, 10$).

für eine komplett richtige Lösung vergeben werden und sonst 0 Punkte.

Die Maximalpunktzahl in der Klausur beträgt wie beim Standardscoring $N := n \cdot k$. Für den Erwartungswert von X gilt wieder:

$$\mathcal{E}(X) = \sum_{i=1}^{n} \mathcal{E}(Y_i) = n \cdot \mathcal{E}(Y),$$

und der Erwartungswert $\mathcal{E}(Y)$ bei einer (m, k)-Aufgabe vereinfacht sich beim AoN-Scoring zu:

$$\begin{aligned} \mathcal{E}(Y) &= k \cdot P(Y = k) \\ &= k \cdot \sum_{w=0}^{k} P(Y = k, W = w) \\ &= k \cdot \sum_{w=0}^{k} P(Y = k \mid W = w) \cdot P(W = w). \end{aligned} \tag{10.7}$$

Dabei ist $P(Y = k \mid W = w)$ wieder durch Gleichung 10.3 definiert und die Zufallsvariable W

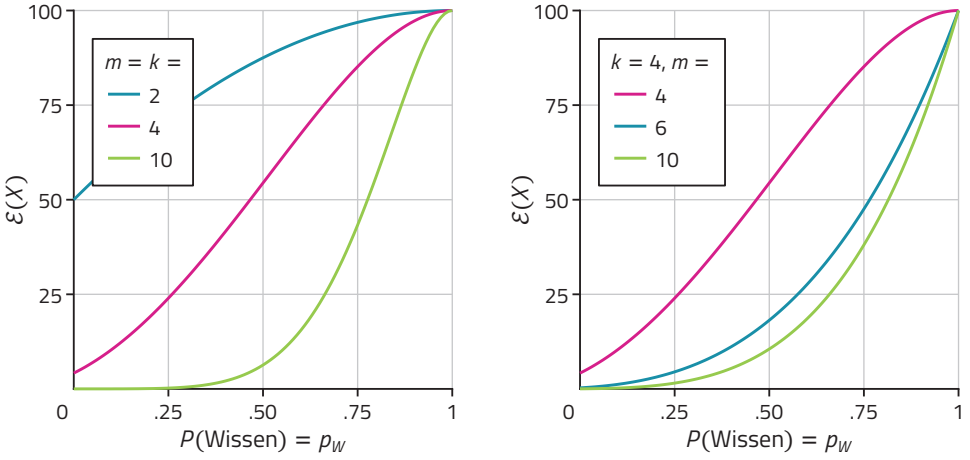


Abbildung 10.5. Erwartungswert für eine Klausur mit n-vielen Aufgaben gleichen Typs als Funktion vom p_W beim AoN-Scoring. Die Anzahl der Aufgaben n wurde jeweils so gewählt, dass immer genau $N = 100$ Punkte maximal in der Klausur zu erreichen waren. Man kann die Ordinate deshalb auch interpretieren als: „Erwartungswert der Punkte in Prozent der Maximalpunktzahl". Links: Erwartungswertfunktionen für eine Klausur mit (k, k)-Zuordnungsaufgaben (von oben nach unten: $k = 2, 4, 10$). Rechts: Erwartungswertfunktionen für eine Klausur mit $(m, 4)$-Zuordnungsaufgaben (von oben nach unten: $m = 4, 6, 10$).

ist unverändert binomialverteilt mit den Parametern k und p_W. Daraus lassen sich wieder die Erwartungswertfunktionen berechnen.

In Abbildung 10.5 sind für das AoN-Scoringverfahren – analog zu Abbildung 10.4 – links die Erwartungswertfunktionen für verschiedene Werte von k in 1-zu-1-Zuordnungsaufgaben dargestellt und rechts die entsprechenden Funktionen für verschiedene Werte von m bei $(m, 4)$-Aufgaben.

Die Kurve für $k = 2$ im linken Bild ist dabei identisch mit der entsprechenden Kurve in Abbildung 10.4, da bei nur zwei Zuordnungen Standardscoring und AoN-Scoring zum selben Ergebnis führen: es kann nur entweder alles richtig oder alles falsch sein. In allen anderen Fällen zeigt sich aber der typische Effekt des AoN-Scorings: die Ratewahrscheinlichkeit wird erheblich reduziert. Dies führt dazu, dass bei $k = 4$ Zuordnungen der Erwartungswert der Punkte das tatsächliche Wissen bereits ziemlich gut abbildet. Wird die Anzahl der Zuordnungen aber weiter erhöht, dann wird die Ratewahrscheinlichkeit „überkompensiert" und es wird zunehmend schwerer, bei lückenhaftem Wissen Punkte zu erhalten. Einen ähnlichen Effekt haben zusätzliche Distraktoren, wie die rechte Seite von Abbildung 10.5 zeigt. In beiden Fällen

sind mögliche Flüchtigkeitsfehler, die sich beim AoN-Scoring besonders gravierend auswirken, noch gar nicht berücksichtigt (s. Abschnitt 7.5.2).

10.5.2 Anordnungsaufgaben

Bei Anordnungsaufgaben ist nicht eindeutig entscheidbar, aus welchen Teilaufgaben sie zusammengesetzt sind (s. Abschnitt 10.4.2). In der Regel wird deshalb eine Anordnungsaufgabe wie eine Einzelaufgabe behandelt, die genau dann richtig beantwortet ist, wenn die Reihenfolge aller Elemente korrekt angegeben wurde. Auch die Zufallsvariable W, das Wissen des Prüflings, ist dementsprechend lediglich eine dichotome Variable, so dass wir es mit dem relativ trivialen Fall einer Einzelaufgabe mit Ratewahrscheinlichkeit $g = 1/k!$ zu tun haben. Dabei ist k die Anzahl der anzuordnenden Elemente.

Werden bei einer k-Anordnungsaufgabe I Punkte bei richtiger Antwort vergeben, dann sind bei einer Klausur mit n-vielen Aufgaben maximal $N := n \cdot I$ Punkte zu erreichen und der Erwartungswert in dieser Klausur ist mit

$$\mathcal{E}(X) = N \cdot \big(g + p_W \cdot (1 - g)\big) \tag{10.8}$$

eine lineare Funktion von p_W mit Ratewahrscheinlichkeit $g = 1/k!$. Zu beachten ist allerdings, dass p_W hier die Wahrscheinlichkeit angibt, die komplette Reihenfolge zu wissen.

In Abbildung 10.6 sind die Erwartungswertfunktionen für verschiedene Werte von k in Anordnungsaufgaben dargestellt. Die Ratewahrscheinlichkeit spielt bereits bei einem $k \geq 4$ so gut wie keine Rolle mehr.

10.6 Bestehens- und Notengrenzen

10.6.1 Zuordnungsaufgaben

Für die Festlegung geeigneter Bestehens- und Notengrenzen geben Abbildung 10.4 und Gleichung 10.4 sehr eindeutige Hinweise. Als Scoringverfahren bei Zuordnungsaufgaben bietet sich, wie bei den meisten anderen Formaten auch, das Standardverfahren mit Ratekorrektur an. Die Problematik des AoN-Scorings ist aus Abbildung 10.5 deutlich erkennbar: Für $(2, 2)$-Aufgaben ist es identisch mit dem Standardverfahren, für $k \geq 2$ wird es mit wachsen-

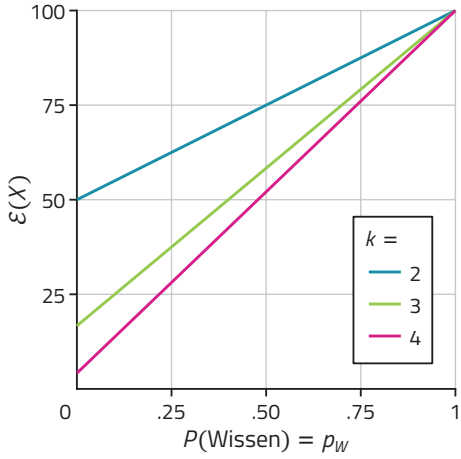


Abbildung 10.6. Erwartungswert für eine Klausur mit n-vielen k-Anordnungsaufgaben als Funktion vom p_W. Die Anzahl der Aufgaben n und die Punktwerte l für korrekte Lösungen wurden jeweils so gewählt, dass genau 100 Punkte maximal in der Klausur zu erreichen waren (von oben nach unten: $k = 2, 3, 4$).

dem k unangemessen schwer, Punkte zu erreichen, wenn man nicht über perfektes Wissen verfügt. Maluspunktverfahren kommen aus den in Abschnitt 10.4.1 genannten Gründen nicht infrage. Beim Standardverfahren besteht das Problem lediglich in der hohen Ratewahrscheinlichkeit für kleine m, die aber in der bekannten Weise durch eine Anpassung der Bestehens- und Notengrenzen berücksichtigt werden kann. Die Formeln für die Berechnung der Ratekorrektur sind zwar nicht mehr linear, aber sie entsprechen gerade den Erwartungswertfunktionen aus Abschnitt 10.5 und können deshalb leicht berechnet werden.

Das Verfahren folgt wieder der in Kapitel 4 skizzierten Vorgehensweise: Wir gehen von einer festgelegten Bewertungsgrenze q in % der Gesamtpunktzahl aus, z. B. „die Note 2.0 erhält, wer mindestens 80% der Maximalpunktzahl erreicht". In Abhängigkeit von der Ratewahrscheinlichkeit bestimmen wir daraus den Erwartungswert für einen Prüfling, der „q% weiß", in unserem Beispiel also den Erwartungswert für einen Prüfling mit $p_W = .80$. Das Ergebnis ist die ratekorrigierte Bewertungsgrenze, die in unserem Beispiel bei einer (4,4)-Zuordnungsaufgabe den Wert 94,76% ergeben würde. Das heißt: Bei (4, 4)-Zuordnungsaufgaben erhält man die Note 2.0 erst ab einem Punktwert, der knapp 95% der Maximalpunktzahl entspricht. Je größer die Ratewahrscheinlichkeit ist, desto höher wird auch die neue Bewertungsgrenze.

In Tabelle 10.2 sind für das an der Martin-Luther-Universität Halle-Wittenberg übliche

Tabelle 10.2. Ratekorrigierte Bestehens- und Notengrenzen für Zuordnungsaufgaben. Am Beispiel des an der Martin-Luther-Universität Halle-Wittenberg üblichen Notenschlüssels werden die korrigierten Grenzen für (m, k)-Zuordnungsaufgaben mit verschiedenen Werten von k und m angegeben. Die Flüchtigkeitsfehlerquote wird mit ($f = 0$) angenommen.

ursprüngliche Notengrenzen in Prozent	50	55	60	65	70	75	80	85	90	95
	ratekorrigierte Notengrenzen (in Prozent)									
(k, k)-Aufgaben										
$k = 2$	87.5	89.9	92.0	93.9	95.5	96.9	98.0	98.9	99.5	99.9
$k = 3$	79.2	82.8	86.1	89.2	91.9	94.3	96.3	97.9	99.0	99.8
$k = 4$	73.4	77.7	81.8	85.5	89.0	92.1	94.8	96.9	98.6	99.6
$k = 5$	69.4	74.0	78.4	82.7	86.6	90.3	93.4	96.1	98.2	99.5
$k = 6$	66.4	71.2	75.9	80.4	84.7	88.7	92.3	95.4	97.8	99.4
$k = 7$	64.2	69.1	73.9	78.6	83.1	87.4	91.3	94.7	97.5	99.3
$k = 8$	62.5	67.4	72.3	77.1	81.8	86.2	90.4	94.1	97.1	99.2
$k = 9$	61.1	66.1	71.0	75.9	80.7	85.3	89.6	93.5	96.8	99.1
$k = 10$	60.0	65.0	69.9	74.9	79.7	84.4	88.9	93.0	96.5	99.0
$(m, 4)$-Aufgaben										
$m = 4$	73.4	77.7	81.8	85.5	89.0	92.1	94.8	96.9	98.6	99.6
$m = 5$	65.3	69.4	73.5	77.4	81.1	84.7	88.2	91.5	94.5	97.4
$m = 6$	61.6	65.8	69.9	74.0	78.0	81.9	85.7	89.5	93.1	96.6
$m = 7$	59.3	63.6	67.9	72.1	76.3	80.4	84.4	88.4	92.4	96.2
$m = 8$	57.8	62.2	66.6	70.9	75.2	79.4	83.6	87.8	91.9	96.0
$m = 9$	56.8	61.2	65.6	70.0	74.4	78.7	83.1	87.3	91.6	95.8
$m = 10$	55.9	60.4	64.9	69.4	73.8	78.3	82.7	87.0	91.4	95.7
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Notensystem die ratekorrigierten Werte für 1-zu-1-Zuordnungen mit $k = 2, \ldots, 10$ und für allgemeine (m, k)-Zuordnungsaufgaben für $k = 4$ und $m = 4, \ldots, 10$ dargestellt. Die Ratewahrscheinlichkeiten bei Zuordnungsaufgaben sind, wie man der Tabelle entnehmen kann, verhältnismäßig hoch und sie steigen vor allem mit zunehmendem Wissen zusätzlich an. Die Ratekorrekturen sind deshalb auch bei größeren Werten von k und m noch beträchtlich, so dass auf sie nicht verzichtet werden kann.

10.6.2 Anordnungsaufgaben

Bei Anordnungsaufgaben wird generell ein Alles-oder-nichts-Scoring verwendet, das typischerweise zu so geringen Ratewahrscheinlichkeiten führt, dass eine Ratekorrektur der Bestehens- und Notengrenzen nur für kleine k nötig ist. Die Ratekorrektur erfolgt nach Gleichung 10.8 durch:

$$q' = 100 \cdot \big(g + q \cdot (1 - g)\big)\,.$$

Dabei ist q die ursprüngliche Notengrenze (in Prozent) und q' die ratekorrigierte Notengrenze. Die Ratewahrscheinlichkeit ist mit $g = 1/k!$ gegeben. Bei einer Anordnung von mehr als vier Elementen kann auf die Ratekorrektur verzichtet werden (s. Abschnitt 10.5.2).

10.7 Zusammenfassung und Schlussfolgerung

Zu- und Anordnungsaufgaben sind einfach zu erstellende Aufgaben und flexibel in der Anzahl der zuzuordnenden Items. Sie erlauben damit auf ökonomische Weise eine Zusammenfassung mehrerer Abfrageelemente. Dazu kommt, dass sie intuitiv leicht zu verstehen sind und nicht zuletzt durch ihre teils spielerische Umsetzung in elektronischen Tests Abwechslung in eine sonst eintönige Abfolge gleichförmiger Aufgabentypen bringen können.

Aufgrund der Abhängigkeiten zwischen den Teilaufgaben ist die Ausprägung der Ratewahrscheinlichkeit nicht ohne Weiteres intuitiv abschätzbar. Sie hängt nicht-linear von der Anzahl der Begriffe und Terme ab – und wird in der Regel vermutlich eher unterschätzt. Umso wichtiger ist eine angemessene Ratekorrektur, deren Berechnung etwas aufwendiger ist als bei den übrigen Aufgabentypen. Da die Auswertung aber in der Regel ohnehin rechnergestützt erfolgt, fällt dieser Nachteil kaum ins Gewicht.

Vor der Verwendung von Zuordnungs- oder Anordnungsaufgaben sollte überlegt werden, ob das zu prüfende Wissen nicht transparenter und eindeutiger mit *single*- und *multiple-response*-Aufgaben abgefragt werden kann. Insbesondere bei Anordnungsaufgaben ist die Beschränkung auf ein Alles-oder-nichts-Scoring in vielen Fällen eine unnötige Einschränkung.

Zur Abfrage von Wissen außerhalb von Prüfungen mit ihren weitreichenden Konsequenzen für die Bewertung von Studierenden, z. B. in Selbsttests oder Lernkontroll-Tests, ist die Verwendung von Zu- und Anordnungsaufgaben oft besonders gut geeignet. Wegen der auf-

lockernden, interaktiven Aufgabenform sind sie bei Studierenden beliebt. Inhaltlich lassen sich mit Zuordnungsaufgaben schnell Beziehungen, Gemeinsamkeiten oder Unterschiede verschiedener Begriffe abfragen, z. B.:

- die Zuordnung von Definitionen zu gegebenen Begriffen,
- die Zuordnung von Abbildungen zu Beschriftungen oder Konzepten, die dargestellt werden,
- die Zuordnung von Handlungsprinzipien zu Anwendungsfällen

und vieles mehr. Mit Anordnungsaufgaben lassen sich typischerweise hierarchische Zusammenhänge und z. B. zeitliche Abfolgen anschaulich erfragen.

Bei der Auswahl der einander zuzuordnenden Begriffe sollte man im Übrigen besonders sorgfältig vorgehen. So sollten alle möglichen Zuordnungen grammatikalisch und logisch möglich sein und alle Definitionen und alle Terme jeweils aus einer Kategorie stammen. Andernfalls können Prüflinge aufgrund von Inkompatibilitäten falsche Zuordnungen ausschließen und damit ihre Ratewahrscheinlichkeit deutlich erhöhen.

11 Aufgaben mit freiem Format

11.1 Charakteristik

In ILIAS gibt es die Möglichkeit, eine Reihe von freien Formaten zu nutzen. Hierbei findet ein Teil des Prüfungsvorgangs außerhalb von ILIAS statt, z. B. das Anfertigen der Antwort oder die spezifische Formulierung der Aufgabenstellung und nur das Resultat wird an die Prüfungsplattform zur Bewertung übertragen. Prüfende erlangen durch dieses Verfahren die Möglichkeit, auch Aufgaben zu stellen, die von den standardmäßig von ILIAS angebotenen Aufgabentypen abweichen oder über diese hinausgehen. Eine automatische Bewertung durch ILIAS ist bei den freien Formaten nicht möglich.

Aufgaben im freien Format können sehr unterschiedliche Formen annehmen und auch sehr unterschiedliche Anforderungen an die Prüflinge stellen. Ein Großteil dieser Aufgaben lässt sich vermutlich in das bereits in Kapitel 9 besprochene Schema der offenen Aufgaben einordnen. Dies ist immer dann der Fall, wenn die Prüflinge ihre Antworten selbst formulieren müssen, ohne dass Sie die Möglichkeit haben, zu raten (z. B., wenn ein Essay oder eine Hausarbeit per Datei-Upload abzugeben ist). Dann gelten die Ausführungen in Kapitel 9 entsprechend.

Die freie Gestaltung der Aufgaben bietet auch die Möglichkeit, andere bereits besprochene Formate nachzustellen. So kann die Aufgabe an die Prüflinge z. B. lauten, bei einer Zeichenaufgabe in einem vorgegebenen Bild bestimmte Bereiche zu markieren. Es handelt sich dann – je nachdem, ob nur ein oder mehrere Bereich in Frage kommen – um eine *single-response*-Aufgabe (s. Kapitel 6) oder um eine *multiple-select*-Aufgabe (s. Kapitel 7). Weiterhin ließe sich mit Hilfe einer *Flash*- oder *Java-Applet*-Frage das bisher nicht in ILIAS zur Verfügung stehende *multiple-true-false*-Format (s. Kapitel 8) implementieren.

Welcher Aufgabentyp auch immer mittels einer Aufgabe im freien Format umgesetzt wird, es ist stets Aufgabe des Prüfenden, einzuschätzen, um welches Format es sich handelt und für die Auswertung das entsprechende Kapitel dieses Handbuchs zu Rate zu ziehen.

11.2 Das freie Aufgabenformat in ILIAS

In ILIAS sind die folgenden Aufgabentypen dem freien Aufgabenformat zuzuordnen:

- Datei hochladen
- Zeichenaufgabe
- *Flash*-Frage
- *Java-Applet*-Frage

11.2.1 Datei hochladen

Beim Aufgabentyp „Datei hochladen“ werden die Prüflinge beauftragt, eine Datei außerhalb von ILIAS zu erstellen, etwa eine Zeichnung, eine Präsentation, Programmcode, einen Aufsatz etc. Diese Datei soll anschließend in ILIAS hochgeladen werden, so dass sie den Prüfenden zur Korrektur zur Verfügung steht. Wie sich eine solche Aufgabe für die Prüflinge darstellt, ist fallspezifisch und hängt von der Aufgabenstellung ab.

Informationen zum Erstellen einer Datei-hochladen-Aufgabe stehen in der ILIAS-Dokumentation zur Verfügung. Ein einfaches Beispiel ist in Abbildung 11.1 dargestellt.

11.2.2 Zeichenaufgabe

Bei einer Zeichenaufgabe stehen den Prüflingen ein Stift in mehreren Farben und ein Radierer als virtuelle Werkzeuge zur Verfügung. Mit deren Hilfe können die Prüflinge je nach Aufgabenstellung auf einer freien Fläche oder in einem vorgegebenen Bild etwas darstellen, z. B. Kurvenverläufe in ein Diagramm einzeichnen oder die Bruchstelle eines Knochens in einem Röntgenbild markieren. Dieser Aufgabentyp wurde am @LLZ als Plugin für ILIAS entwickelt (Jobst & Annanias, 2014b). Ein einfaches Beispiel ist in Abbildung 11.2 dargestellt.

Schreiben Sie eine Hausarbeit im Umfang von 20 Seiten über die Systemtheorie nach Niklas Luhmann. Laden Sie anschließend Ihr Dokument im pdf-Format in ILIAS hoch.

BEREITS ABGEGEBENE DATEIEN

Dateiname	Datum
Keine Einträge	

Datei hochladen

Browse... No file selected.

Hochladen

Maximal erlaubte Upload-Größe: 250.0 MB

Abbildung 11.1. Beispiel einer Datei-hochladen-Aufgabe in ILIAS. Es handelt sich um eine offene Aufgabe mit einer Ratewahrscheinlichkeit von $g = 0$ (s. Kapitel 9).

In einem Experiment zur Signalentdeckungstheorie wurden die Daten in der untenstehenden Tabelle erhoben (absolute Häufigkeiten). Tragen Sie in das Diagramm die vier Datenpunkte der ROC-Kurve ein und skizzieren Sie den ungefähren Verlauf der dazu am besten passenden ROC-Kurve.

	sicher ja	eher ja	???	eher nein	sicher nein	Summe
signal	30	45	10	5	10	100
noise	10	10	10	30	40	100

Rückgängig | Wiederholen | Zeichnen | Radierer | Alles Löschen

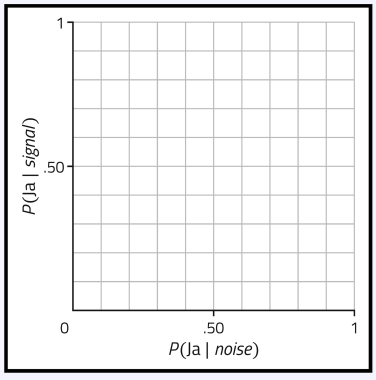


Abbildung 11.2. Beispiel einer Zeichenaufgabe in ILIAS. Es handelt sich um eine offene Aufgabe mit einer Ratewahrscheinlichkeit von $g = 0$ (s. Kapitel 9).

11.2.3 *Flash*-Frage und *Java-Applet*-Frage

Die Aufgaben vom Typ *Flash*-Frage bzw. *Java-Applet*-Frage können sehr vielfältig sein. Für beide Aufgabentypen gilt, dass ein extern erstelltes Programm in die Aufgabe eingebettet wird und von den Prüflingen bearbeitet werden muss. Welcher Natur dieses Programm ist, liegt in der Hand des Prüfenden.

Natürlich lassen sich alle bereits in den vorangegangenen Kapiteln besprochenen Aufgabentypen auch mittels *Flash* oder *Java* nachbilden. Das wäre aber nicht nur unökonomisch, sondern würde zusätzlich eine automatische Bewertung der Antworten durch ILIAS verhindern. Diese beiden Aufgabentypen bieten sich jedoch an, wenn z. B. umfangreiche Prozesse simuliert werden sollen und die Prüflinge an bestimmten kritischen Stellen der Prozesse eine Entscheidung über den Fortgang der Simulation treffen müssen.

Flash-Frage bzw. *Java-Applet*-Frage sind deshalb besonders dann empfehlenswert, wenn Handlungswissen überprüft werden soll, haben jedoch den Nachteil, dass das einzubettende Programm zunächst außerhalb von ILIAS programmiert werden muss.

12 Kombination von mehreren Aufgabentypen

In den vorangegangenen Kapiteln wurde jeweils ein spezifisches Aufgabenformat anhand einer Beispielklausur besprochen, die aus Aufgaben nur dieses einen Typs bestand. Die Beschränkung auf einen einheitlichen Aufgabentyp in einer Klausur wird in der Realität häufig praktiziert und kann gute Gründe haben. Im medizinischen Staatsexamen werden etwa ausschließlich *single-response*-Aufgaben mit jeweils fünf Antwortalternativen verwendet, bei der theoretischen Führerscheinprüfung nur *multiple-select*-Aufgaben. Der größte Vorteil eines einheitlichen Aufgabenformates besteht darin, dass die Prüflinge mit dem Aufgabentyp vertraut sind, bzw. bei der Vorbereitung schnell damit vertraut gemacht werden können und die Prüfenden ein einheitliches, einfaches Auswertungsschema anwenden können. Angesichts der zunehmenden Verbreitung von E-Learning-Plattformen mit ihren vielfältigen Test- und Prüfungsmöglichkeiten kann man allerdings inzwischen wohl davon ausgehen, dass Prüflinge mit den verschiedenen in diesem Handbuch beschriebenen Aufgabentypen hinreichend gut vertraut sind. Auch das Argument der einfacheren Auswertung ist nicht mehr sehr stichhaltig: Die in diesem Handbuch beschriebenen Prinzipien der Ratekorrektur lassen sich problemlos auf Klausuren mit einem bunten Mix aus unterschiedlichen Aufgabentypen anwenden. Da die Auswertung der Klausuren beim *E-Assessment* ohnehin automatisch erfolgt, spielt der höhere Aufwand für Berechnungen keine Rolle. Das Erstellen einer Klausur wird dagegen deutlich einfacher, wenn nicht alle Aufgaben im selben Format formuliert werden müssen, sondern je nach Inhalt und Struktur des zu prüfenden Wissens das dafür am besten geeignete Format gewählt werden kann. Im Folgenden Kapitel wird beschrieben, wie bei der Bestimmung ratekorrigierter Bestehens- und Notengrenzen für eine Klausur mit unterschiedlichen Aufgabentypen vorzugehen ist.

12.1 Einheitliches Format oder Aufgabenmix?

Die Entscheidung für einen bestimmten Aufgabentyp sollte sich ausschließlich an der Eignung für das zu prüfende Wissen orientieren. Wenn es zu einer richtigen Antwort typische Falschantworten gibt und die Diskriminationsfähigkeit geprüft werden soll, bieten sich *single-response*-Aufgaben an. Wenn bei der Führerscheinprüfung gefragt wird, in welcher Reihenfolge die dargestellten Fahrzeuge eine bestimmte Kreuzung passieren dürfen, ist es viel naheliegender, eine Anordnungsaufgabe zu formulieren, als verschiedene Reihenfolgen vorzugeben und die richtige auswählen zu lassen. Für kleinere Rechenaufgaben eignen sich numerische Texteingaben am besten usw. Mit den heute zur Verfügung stehenden technischen Mitteln ist es sehr einfach geworden, verschiedene Aufgabentypen und -formate in einer Klausur zu verwenden. Das ermöglicht eine sehr flexible Gestaltung von treffsicheren Klausuren ohne den Zwang zu einem einheitlichen Format.

Auf der anderen Seite sollte die Vielfalt der Formate aber nicht zum Selbstzweck werden. Die Verwendung von vielen verschiedenen Aufgabentypen in einer Klausur stellt auch Anforderungen an die Flexibilität der Studierenden und die Empfehlung lautet auch hier: Verwenden Sie so viele Aufgabentypen wie nötig, aber so wenige wie möglich. Auf jeden Fall muss bei der Vorbereitung auf die Klausur sichergestellt sein, dass die Studierenden mit allen verwendeten Aufgabentypen hinreichend gut vertraut sind und wissen, was genau von ihnen beim Beantworten erwartet wird. Insbesondere muss bei jeder einzelnen Aufgabe der Aufgabentyp klar erkennbar sein.

12.2 Bestehens- und Notengrenzen

Die Berechnung ratekorrigierter Bestehens- und Notengrenzen basiert auf der Berechnung von Erwartungswerten für die Zufallsvariable X (Anzahl der erreichten Punkte) in Abhängigkeit vom Wissen p_W. Für p_W sind die ursprünglichen – unkorrigierten – Bestehens- und Notengrenzen einzusetzen (vgl. Kapitel 4). Da für die Gesamtklausur die bei jeder Aufgabe erzielten Punkte summiert werden und der Erwartungswert der Summe gleich der Summe der Erwartungswerte jeder einzelnen Aufgabe ist, genügt es

- für jede Aufgabe i jeweils den Erwartungswert $E(X_i)$ zu berechnen und
- für die Gesamtklausur die Summe dieser Erwartungswerte zu bilden.

Die Erwartungswerte der einzelnen Aufgaben werden entsprechend ihrer Ratewahrscheinlichkeit und ggf. eines Wertes für die Flüchtigkeitsfehlertoleranz so berechnet, wie dies in den entsprechenden Kapiteln beschrieben ist. Dabei wird immer das Standard-Scoring zugrundegelegt, mit Ausnahme der Anordnungsaufgaben, bei denen es keine Alternative zum Alles-oder-nichts-Scoring gibt. Beim *formula scoring* und beim *testlet scoring* ergibt eine Ratekorrektur keinen Sinn, da dem Fehler erster Art hier mit anderen Methoden begegnet wird. Die „Rateneigung" der Prüflinge wird aus den in Kapitel 4 genannten Gründen konstant mit dem Wert $h = 1$ angenommen.

12.3 Beispiel

Für jede Klausur wird eine Tabelle nach dem Muster von Tabelle 12.1 angelegt, in der alle Aufgaben mit ihren spezifischen Parametern aufgeführt werden. Die Beispielklausur in Tabelle 12.1 besteht aus insgesamt 22 Aufgaben: fünf *single-response*-, fünf *multiple-select*-, drei *multiple-true-false*-, vier offenen, zwei Anordnungs- und drei Zuordnungsaufgaben. Insgesamt können maximal 100 Punkte erreicht werden. Für jede einzelne Aufgabe sind im linken Teil der Tabelle die Aufgabenparameter und die Punktbewertung angegeben. Im rechten Teil der Tabelle werden daraus die ratekorrigierten Bestehens- und Notengrenzen pro Zeile für jede Aufgabe berechnet. Der zugrundegelegte Original-Notenschlüssel ist rechts oben angegeben. Im Beispiel in Tabelle 12.1 sind das Schulnoten (1 ≙ „sehr gut", 2 ≙ „gut", ..., 5 ≙ „mangelhaft" und 6 ≙ „ungenügend") mit der Annahme, dass für die Note 1 mindestens 90% des Stoffes beherrscht werden müssen, für die Note 2 mindestens 80% usw.

Die ratekorrigierten Notengrenzen für die Gesamtklausur stehen ganz unten in der letzte Zeile und ergeben sich als Summen der Spalten. Bei dieser Klausur müssen für die Note 4 mindestens 70.1 Punkte (70.1% von 100 Punkten) erreicht werden, für die Note 3 mindestens 77.8% und für die Note 1 braucht man mindestens 92.7% der maximalen Punktzahl. Wegen der Ratewahrscheinlichkeiten liegen die Notengrenzen immer etwas höher als beim Original-Notenschlüssel. Bei einer anderen Klausur mit anderen Ratewahrscheinlichkeiten ändern sich die Notengrenzen entsprechend.

Die beiden Spalten für $p_W = 0$ und $p_W = 1$ werden für die Berechnung von Notengrenzen im Allgemeinen nicht benötigt. Sie sollen lediglich demonstrieren, wie groß der Erwartungswert für die erreichte Punktzahl in den beiden Extremfällen „zufälliges Raten bei jeder einzelnen

Aufgabe“ und „lückenloses Wissen bei allen Aufgaben“ ist und die Tabelle damit transparenter machen.

Die Berechnung der ratekorrigierten Bestehens- und Notengrenzen erfordert viel weniger Aufwand, als es auf den ersten Blick erscheinen mag. Sowohl die Auflistung der verwendeten Aufgaben und ihrer Parameter als auch die eigentliche Berechnung der Notengrenzen kann problemlos von der Software übernommen werden, mit der die Aufgaben erstellt werden. Für die Prüfungssoftware EvaExam (Electric Paper Evaluationssysteme GmbH, 2017a) gibt es z. B. ab der Version 7.1 ein Plug-In (*Anpassungstool für Notenschlüssel*) zur automatischen Berechnung der ratekorrigierten Bestehens- und Notengrenzen für beliebige Notenschlüssel. In ILIAS wird das vermutlich bald ebenfalls der Fall sein.

Falls die Aufgaben manuell erstellt werden oder mit einer Software, die noch keine Ratekorrekturen unterstützt, kann die Liste der Aufgaben und ihrer Parameter auch gut manuell zusammengestellt werden. Für die Berechnung der ratekorrigierten Notengrenzen (rechter Teil von Tabelle 12.1) ist dafür auf der Internetpräsenz des @LLZ ein Berechnungstool mit den wichtigsten Formeln verfügbar. Werden in dieses Tool die einzelnen Aufgaben mit ihren Parametern (linker Teil von Tabelle 12.1) und ein Notenschlüssel (Tabelle 12.1 rechts oben) eingetragen, erfolgt automatisch die Berechnung der Notengrenzen.

Für Klausuren mit nur wenigen unterschiedlichen Aufgabentypen ist es oft einfacher, für jeden Aufgabentyp den entsprechenden Parametersatz nur einmal aufzuführen und die Anzahl der Aufgaben dieses Typs anzugeben. Ein Beispiel dafür ist in Tabelle 12.2 angegeben.

Tabelle 12.1. Berechnung der ratekorrigierten Notengrenzen für eine Beispiel-Klausur mit 22 Aufgaben unterschiedlichen Typs. Am Beispiel eines fiktiven Schulnotenschlüssels werden die korrigierten Grenzen für die Gesamtklausur als Summe der Grenzen für jede einzelne Aufgabe berechnet.

							Schulnoten						
							6	5	4	3	2	1	
							unkorrigierte Notengrenzen ≙ p_W						
							0	.50	.60	.70	.80	.90	1
		Parameter											
Nr.	Typ	*k*	*m*	*f*	*g*	Punkte	ratekorrigierte Notengrenzen ≙ $E(p_W)$						
1	SR		5	.05	.20	1	.20	.58	.65	.73	.80	.88	.95
2	SR		3	.05	.30	1	.33	.64	.70	.77	.83	.89	.95
3	SR		4	.05	.25	1	.25	.60	.67	.74	.81	.88	.95
4	SR		5	.05	.20	1	.20	.58	.65	.73	.80	.88	.95
5	SR		4	.05	.25	1	.25	.60	.67	.74	.81	.88	.95
6	MS	4	2	.01	.50	4	2.00	2.98	3.18	3.37	3.57	3.76	3.96
7	MS	6	2	.01	.50	6	3.00	4.47	4.76	5.06	5.35	5.65	5.94
8	MS	5	2	.01	.50	5	2.50	3.73	3.97	4.22	4.46	4.71	4.95
9	MS	4	2	.01	.50	4	2.00	2.98	3.18	3.37	3.57	3.76	3.96
10	MS	4	2	.01	.50	4	2.00	2.98	3.18	3.37	3.57	3.76	3.96
11	MTF	4	2	0	.50	4	2.00	3.00	3.20	3.40	3.60	3.80	4.00
12	MTF	6	2	0	.50	6	3.00	4.50	4.80	5.10	5.40	5.70	6.00
13	MTF	4	2	0	.50	4	2.00	3.00	3.20	3.40	3.60	3.80	4.00
14	OF			0	0	10	0	5.00	6.00	7.00	8.00	9.00	10.0
15	OF			0	0	4	0	2.00	2.40	2.80	3.20	3.60	4.00
16	OF			0	0	8	0	4.00	4.80	5.60	6.40	7.20	8.00
17	OF			0	0	10	0	5.00	6.00	7.00	8.00	9.00	10.0
18	AN	4		0	.04	4	.17	2.08	2.47	2.85	3.23	3.62	4.00
19	AN	6		0	0	6	.01	3.00	3.60	4.20	4.80	5.40	6.00
20	ZU	4	5	0	dyn.	4	.80	2.61	2.94	3.25	3.53	3.78	4.00
21	ZU	6	6	0	dyn.	6	1.00	3.98	4.55	5.08	5.54	5.87	6.00
22	ZU	6	6	0	dyn.	6	1.00	3.98	4.55	5.08	5.54	5.87	6.00
					Summe	**100**	**22.7**	**62.3**	**70.1**	**77.8**	**85.4**	**92.7**	**99.5**

Tabelle 12.2. Berechnung der ratekorrigierten Notengrenzen für eine Beispiel-Klausur mit 18 *single-response*-, fünf *multiple-true-false*- und zehn offenen Aufgaben. Die *single-response*-Aufgaben mit unterschiedlicher Ratewahrscheinlichkeit werden getrennt aufgelistet. Alle offenen Fragen können dagegen zusammengefasst werden, anzugeben ist lediglich die Summe der erreichbaren Punkte.

								Schulnoten						
								6	5	4	3	2	1	
								unkorrigierte Notengrenzen ≙ p_W						
								0	.50	.60	.70	.80	.90	1
		Parameter												
Anz.	Typ	k	m	f	g	Punkte	Σ	ratekorrigierte Notengrenzen ≙ $E(p_W)$						
13	SR		5	.05	.20	1	13	2.60	7.48	8.45	9.43	10.4	11.4	12.3
5	SR		4	.05	.25	1	5	1.25	3.00	3.35	3.70	4.05	4.40	4.75
5	MTF	4	2	0	.50	4	20	10.0	15.0	16.0	17.0	18.0	19.0	20.0
10	OF			0	0		62	0	31.0	37.2	43.4	49.6	55.8	62.0
						Summe	**100**	**13.8**	**56.5**	**65.0**	**73.5**	**82.0**	**90.6**	**99.1**

Teil III

Kurzreferenz zu den Aufgabenformaten

13 Cheat Sheets

Dieser Teil des Handbuchs lässt sich als schnelles Nachschlagewerk über die Erkenntnisse verstehen, die durch die Anwendung des Wahrscheinlichkeitsmodells aus Teil I auf die in Teil II besprochenen Aufgabenformate, gewonnen wurden. Wer die ersten beiden Teile gelesen hat, findet hier eine knappe Zusammenfassung der Inhalte. Prüfende, die nach einer theoretisch fundierten Möglichkeit zur Prüfungsbewertung suchen, sich jedoch nicht mit den zugegebenermaßen nicht immer leicht verständlichen mathematischen Hintergründen auseinandersetzen möchten, finden hier für die einzelnen Aufgabentypen genaue Handlungsanweisungen

- zur Bepunktung der Aufgaben, dem sogenannten Scoring (genaueres dazu s. Kapitel 3),
- zum Verlauf der zu erwartenden Punkte bei unterschiedlich großem Wissen (genaueres dazu s. die einzelnen Kapitel in Teil II) und
- zur Ratekorrektur von Bestehens- und Notengrenzen (genaueres dazu s. Kapitel 4).

An dieser Stelle sei darauf hingewiesen, dass bei den im Folgenden dargestellten Handlungsanweisungen zur Prüfungsauswertung stets davon ausgegangen wird, dass die Prüflinge die Freiheit besitzen, zu raten. Die dadurch ggf. durch „Glück" erratenen Punkte werden bei der Berechnung der Bestehens- und Notengrenzen berücksichtigt. Prüflinge, die bei Nichtwissen nicht raten, werden durch diese Herangehensweise gegenüber ratenden Prüflingen systematisch benachteiligt. Es ist daher empfehlenswert, im Vorfeld der Prüfung

- das Bewertungsschema bekanntzugeben und
- die Studierenden aufzufordern, auf jeden Fall alle Aufgaben zu beantworten, keine Aufgaben auszulassen und im Zweifel die Antwortalternativen anzukreuzen, die am ehesten in Frage kommen.

Für die Abbildungen und Tabellen auf den folgenden Seiten wird daher für die Rateneigung stets der Wert $h = 1$ angenommen. Weiterhin wurde für Flüchtigkeitsfehler das strikteste Kriterium angelegt ($f = 0$), so dass die Darstellungen als Obergrenzen zu verstehen sind. Wird ein

milderes Kriterium angelegt, verringern sich die Bestehens- und Notengrenzen entsprechend.

Falls die Aufgaben manuell erstellt werden oder mit einer Software, die noch keine Ratekorrekturen unterstützt, kann die Liste der Aufgaben und ihrer Parameter auch gut manuell zusammengestellt werden. Für die Berechnung der ratekorrigierten Notengrenzen ist dafür auf der Internetpräsenz des @LLZ ein Berechnungstool mit den wichtigsten Formeln verfügbar. Werden in dieses Tool die einzelnen Aufgaben mit ihren Parametern und ein Notenschlüssel eingetragen, erfolgt automatisch die Berechnung der Notengrenzen.

Cheat Sheet zum *single-response*-Format

Zugehörige ILIAS-Aufgabentypen

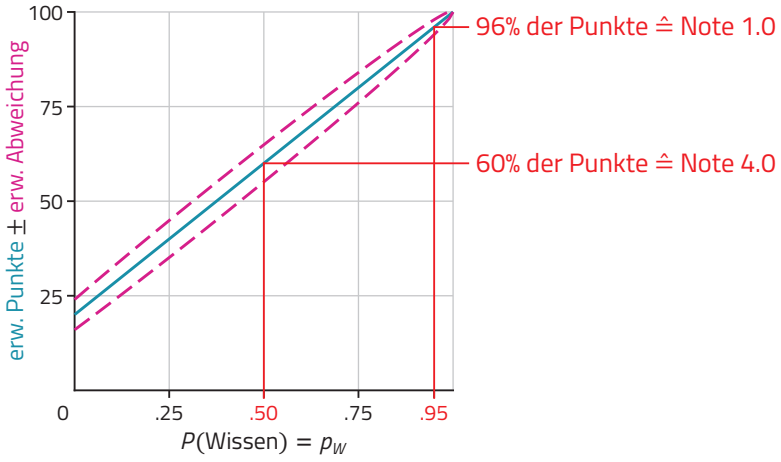
- *ImageMap* mit dem Antwortmodus „*Single Choice*“
- Lückentext vom Typ Auswahl-Lücke
- *Single Choice*

Punktevergabe

- wenn ausschließlich die tatsächlich richtige Alternative angekreuzt ist: **1 Punkt** (Gewichtung möglich)
- wenn eine falsche Alternative angekreuzt ist: **0 Punkte**
- wenn mehrere Alternativen (auch die richtige) angekreuzt sind: **0 Punkte**
- wenn keine Alternative angekreuzt ist: **0 Punkte**

Zusammenhang zwischen Wissen und zu erwartenden Punkten

Bei diesem Verlauf der zu erwartenden Punkte handelt es sich um ein Beispiel für eine Aufgabe mit fünf Alternativen ($g = .20$).



Bestehens- und Notengrenzen

% Wissen	50	55	60	65	70	75	80	85	90	95
	zu erwartende Punkte in Prozent									
bei 3 Alternativen	66.7	70.0	73.3	76.7	80.0	83.3	86.7	90.0	93.3	96.7
bei 4 Alternativen	62.5	66.3	70.0	73.8	77.5	81.3	85.0	88.8	92.5	96.3
bei 5 Alternativen	60.0	64.0	68.0	72.0	76.0	80.0	84.0	88.0	92.0	96.0
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Cheat Sheet zum *multiple-response*-Format

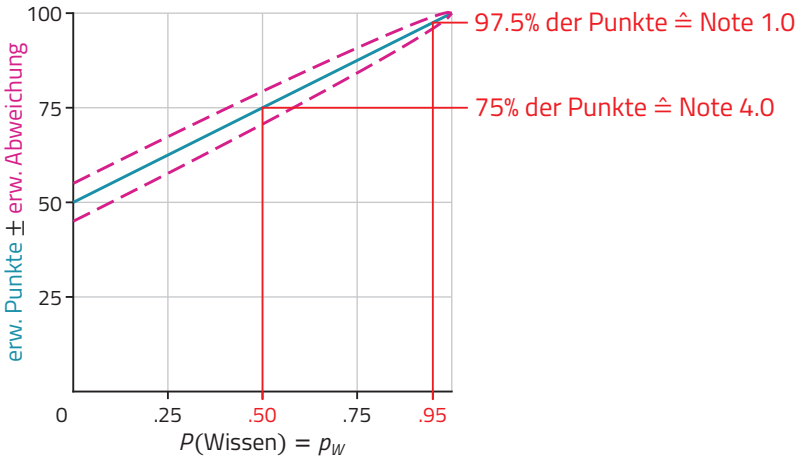
Zugehörige ILIAS-Aufgabentypen

- *Multiple Choice*
- *ImageMap* mit dem Antwortmodus „*Multiple Choice*“
- Fehlertext

Punktevergabe

- jede (als richtig) angekreuzte tatsächlich richtige Alternative: **1 Punkt** (Gewichtung möglich)
- jede nicht angekreuzte oder als falsch angekreuzte tatsächlich falsche Alternative: **1 Punkt** (Gewichtung möglich)
- ansonsten: **0 Punkte**

Zusammenhang zwischen Wissen und zu erwartenden Punkten



Bestehens- und Notengrenzen

% Wissen	50	55	60	65	70	75	80	85	90	95
	zu erwartende Punkte in Prozent									
Einzelbewertung	75.0	77.5	80.0	82.5	85.0	87.5	90.0	92.5	95.0	97.5
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Cheat Sheet zu Zuordnungsaufgaben

Zugehörige ILIAS-Aufgabentypen

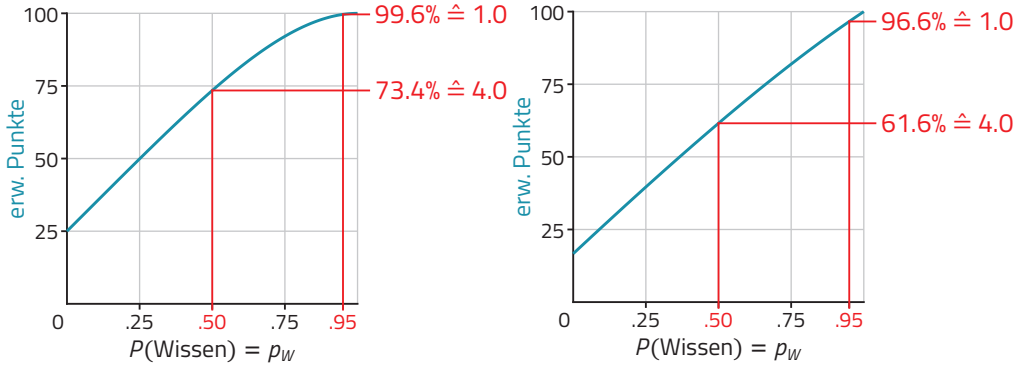
- Zuordnungsfrage

Punktevergabe

- für jede korrekte Zuordnung: **1 Punkt** (Gewichtung möglich)
- ansonsten: **0 Punkte**

Zusammenhang zwischen Wissen und zu erwartenden Punkten

Der Verlauf der zu erwartenden Punkte ist abhängig von der Anzahl der Definitionen (k) und der zuzuordnenden Terme (m). Zuordnungsaufgaben unterscheiden sich danach, ob Distraktoren erlaubt sind ((m, k)-Aufgaben) oder nicht (1-zu-1-Zuordnungen bzw. (k, k)-Aufgaben).
Der Verlauf der zu erwartenden Punkte ist hier beispielhaft für eine $(4, 4)$-Aufgabe (links) und eine $(6, 4)$-Aufgabe (rechts) dargestellt.



Bestehens- und Notengrenzen

% Wissen	50	55	60	65	70	75	80	85	90	95
	zu erwartende Punkte in Prozent									
(k, k)-Aufgaben										
$k = 3$	79.2	82.8	86.1	89.2	91.9	94.3	96.3	97.9	99.0	99.8
$k = 4$	73.4	77.7	81.8	85.5	89.0	92.1	94.8	96.9	98.6	99.6
$k = 5$	69.4	74.0	78.4	82.7	86.6	90.3	93.4	96.1	98.2	99.5
$(m, 4)$-Aufgaben										
$m = 5$	65.3	69.4	73.5	77.4	81.1	84.7	88.2	91.5	94.5	97.4
$m = 6$	61.6	65.8	69.9	74.0	78.0	81.9	85.7	89.5	93.1	96.6
$m = 7$	59.3	63.6	67.9	72.1	76.3	80.4	84.4	88.4	92.4	96.2
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Cheat Sheet zu Anordnungsaufgaben

Zugehörige ILIAS-Aufgabentypen

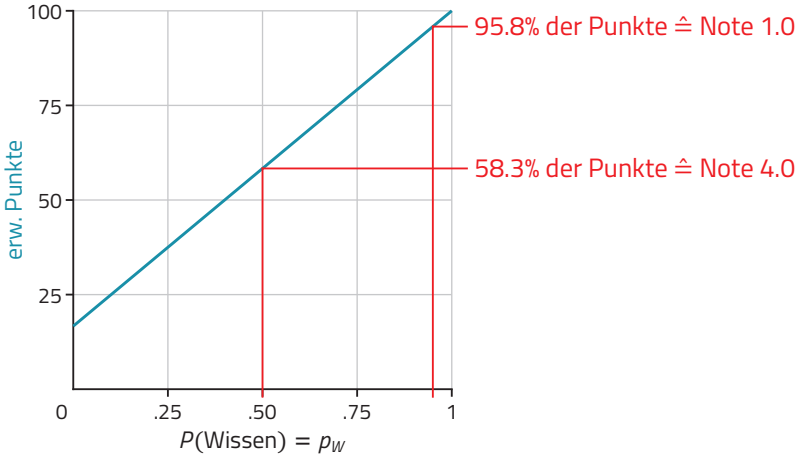
- Anordnungsfrage
- Anordnungsfrage (horizontal)

Punktevergabe

- wenn die gesamte Anordnung korrekt ist (Alles-oder-nichts-Scoring): **1 Punkt** (Gewichtung möglich)
- ansonsten: **0 Punkte**

Zusammenhang zwischen Wissen und zu erwartenden Punkten

Bei diesem Verlauf der zu erwartenden Punkte handelt es sich um ein Beispiel für eine Aufgabe mit drei anzuordnenden Elementen.



Bestehens- und Notengrenzen

Bei einer Anordnung von mehr als vier Elementen kann auf die Ratekorrektur verzichtet werden (s. Abschnitt 10.5.2).

% Wissen	50	55	60	65	70	75	80	85	90	95
	zu erwartende Punkte in Prozent									
bei 2 Elementen	75.0	77.5	80.0	82.5	85.0	87.5	90.0	92.5	95.0	97.5
bei 3 Elementen	58.3	62.5	66.7	70.8	75.0	79.2	83.3	87.5	91.7	95.8
bei 4 Elementen	52.1	56.9	61.7	66.5	71.2	76.0	80.8	85.6	90.4	95.2
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Cheat Sheet zu offenen Aufgaben

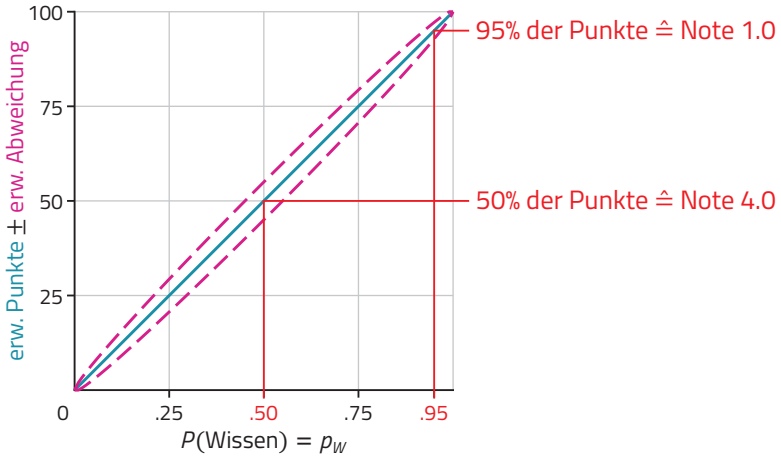
Zugehörige ILIAS-Aufgabentypen

- Formelfrage
- Freitext
- *JSME*-Frage
- Lückentext (Numerische Lücke)
- Lückentext (Textlücke)
- *Long-Menu*-Frage
- Numerische Frage
- Text-Teilmenge

Punktevergabe

- jedes richtige Teilergebnis: **1 Punkt** (Gewichtung möglich).
  Teilergebnisse können sein:
  - bei Formelfragen und numerischen Fragen: die einzelnen Ergebnisse
  - bei Lückentexten: die Lösungen jeder einzelnen Lücke
  - bei Freitexten und Text-Teilmengen: die erwarteten Begriffe
  - bei *JSME*-Fragen: Strukturen bzw. Teilstrukturen
- ansonsten: **0 Punkte**.

Zusammenhang zwischen Wissen und zu erwartenden Punkten



Bestehens- und Notengrenzen

% Wissen	50	55	60	65	70	75	80	85	90	95
	zu erwartende Punkte in Prozent									
Einzelbewertung	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0	95.0
Note	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

Cheat Sheet zu Aufgaben mit freiem Format

Zugehörige ILIAS-Aufgabentypen

- Datei hochladen
- *Flash*-Frage
- *Java-Applet*-Frage
- Zeichenaufgabe

Punktevergabe

Beim freien Antwortformat wird die Qualität der Antwort vom Prüfenden mit einem Punktwert zwischen 0 und *n*, der maximalen Punktzahl für diese Aufgabe, bewertet. Dabei kann es sinnvoll sein, je nach den Anforderungen der Aufgabe vor der Bewertung die Kriterien für die Vergabe von Punkten festzulegen.

Zusammenhang zwischen Wissen und zu erwartenden Punkten

Da den Prüfenden bei der Gestaltung einer Aufgabe mit freiem Format keinerlei Grenzen gesetzt sind, kann an dieser Stelle kein allgemeiner Zusammenhang zwischen Wissen und zu erwartenden Punkten formuliert werden. Dies liegt daran, dass die wichtige Größe der Ratewahrscheinlichkeit *g* (s. Kapitel 2) für die verschiedenen hier denkbaren Aufgaben stark variieren kann. Hierzu einige Beispiele:

- mittels Datei hochladen einen Aufsatz schreiben lassen: $g = 0$ (s. Seite 125)
- mittels *Flash*-Frage *multiple-select*-Aufgaben nachbilden: $g = .50$ (s. Seite 122)
- mittels Zeichenaufgabe eine Funktion skizzieren lassen: $g = 0$ (s. Seite 125)

Bestehens- und Notengrenzen

Eine eventuell erforderliche Ratekorrektur der Bestehens- und Notengrenzen ist entsprechend der Ratewahrscheinlichkeit so durchzuführen, wie dies bei den anderen Formaten beschrieben ist.

Teil IV

Anhang

Abbildungen

Tabellen

Literatur

ÄApprO. (2013). 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. Zugriff unter http://www.gesetze-im-internet.de/_appro_2002/BJNR240500002.html

Bar-Hillel, M., Budescu, D. V. & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, *4*, 3–12. doi:10.1007/s11299-005-0001-z

Bienfait, B. & Ertl, P. (2013). JSME: A free molecule editor in JavaScript. *Journal of Cheminformatics*, *5*, 24. Journal Article. doi:10.1186/1758-2946-5-24

Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika*, *26*, 255–280. doi:10.1007/BF02289796

Brüstle, P. (2011). *Kurzanleitung Prüfen mit MC-Fragen*. Universität Freiburg. Zugriff unter https://www.medizinstudium.uni-freiburg.de/lehrende/pruefungen/kurzanleitung-pruefen-mit-mc-fragen.pdf

Budescu, D. V. & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, *80*, 1105–1122. doi:10.1007/s11336-014-9425-x

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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

Diekert, V., Kufleitner, M. & Rosenberger, G. (2013). *Elemente der diskreten Mathematik: Zahlen und Zählen, Graphen und Verbände*. Berlin: De Gruyter.

ELAN e.V. (2016). E-Assessments & E-Klausuren: E-Prüfungen an Hochschulen. Zugriff unter http://ep.elan-ev.de/wiki/

Electric Paper Evaluationssysteme GmbH. (2017a). EvaExam (Version 7.1).

Electric Paper Evaluationssysteme GmbH. (2017b). EvaSys (Version 7.1).

Endemol Deutschland GmbH. (2016). Wer wird Millionär? [Fernsehsendung]. Deutschland: RTL Television.

GoeEle@rn. (2017). ILIAS 5.1.4 – Neue Features #2: Die Kprim Choice Frage. Zugriff unter http://blog.stud.uni-goettingen.de/goeelearn/2016/05/24/ilias-5-1-4-neue-features-2-die-kprim-choice-frage

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.

Hanson, D., Seyffarth, K. & Weston, J. H. (1983). Matchings, Derangements, Rencontres. *Mathematics Magazine*, *56*, 224–229. doi:10.2307/2689812

Holzinger, K. J. (1924). On scoring multiple response tests. *Journal of Educational Psychology*, *15*, 445–447. doi:10.1037/h0073083

ILIAS open source e-Learning e.V. (2017). ILIAS (Version 5.1.18 2017-05-17).

Jacobs, K. & Jungnickel, D. (2004). *Einführung in die Kombinatorik* (2. Aufl.). Berlin: De Gruyter.

Jančařík, A. & Kostelecká, Y. (2015). The scoring of matching questions tests: A closer look. *The Electronic Journal of e-Learning*, *13*(4), 268–276. Zugriff unter http://www.ejel.org/volume13/issue4/p268

Jobst, C. & Annanias, Y. (2014a, 8. Dezember). Moleküleditoraufgabentyp. Aachen, Deutschland. Zugriff unter http://www.e-pruefungs-symposium.de/wp-content/uploads/2016/03/Abstractband.pdf

Jobst, C. & Annanias, Y. (2014b, 8. Dezember). Zeichenaufgabentyp für ILIAS. Aachen, Deutschland. Zugriff unter http://www.e-pruefungs-symposium.de/wp-content/uploads/2016/03/Abstractband.pdf

Klein, M. (2016). Klausurerfolg trotz Unwissenheit: Akademische Leistungsbewertung mit dem Antwort-Wahl-Verfahren. *Forschung & Lehre*, *23*(1), 38–39. Zugriff unter http://www.forschung-und-lehre.de/wordpress/?p=20114

Krebs, R. (2004). *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung*. Universität Bern. Zugriff unter https://www.ectaveo.ch/Mediathek/2015/08/Anleitung_MC-Fragen-Uni-Bern.pdf

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

Lesage, E., Valcke, M. & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education - Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, *39*, 188–193. doi:10.1016/j.stueduc.2013.07.001

Lieberman, D. A. (2011). *Human learning and memory*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139046978

Lindner, M. A., Strobel, B. & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung. *Zeitschrift für Pädagogische Psychologie*, *29*, 133–149. doi:10.1024/1010-0652/a000156

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. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M., Novick, M. R. & Birnbaum, A. (2008). *Statistical theories of mental test scores*. Charlotte, NC: Information Age Publishing.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (Hrsg.), *Handbook of mathematical psychology* (S. 1–103). Oxford, UK: John Wiley & Sons.

Luce, R. D. & Galanter, E. (1963). Discrimination. In R. D. Luce, R. R. Bush & E. Galanter (Hrsg.), *Handbook of mathematical psychology* (S. 191–243). Oxford, UK: John Wiley & Sons.

Ludwig, J. (2014, 8. Oktober). Wenn nichts mehr geht. *Zeit Campus*. Zugriff unter http://www.zeit.de/campus/2014/06/pruefungsergebnis-klage

Melzer, A. (2016). *Auswertungsverfahren für Prüfungen mit* multiple response*-Aufgaben auf der Grundlage der Signalentdeckungstheorie*. Dissertationsschrift, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale). Zugriff unter http://nbn-resolving.de/urn:nbn:de:gbv:3:4-18601

OVG Nordrhein-Westfalen, Urteil vom 16. Dezember 2008. Aktenzeichen 14 A 2154/08. Zugriff unter https://openjur.de/u/134912.print

OVG Nordrhein-Westfalen, Urteil vom 21. Juni 2016. Aktenzeichen 14 A 3066/15. Zugriff unter http://www.justiz.nrw.de/nrwe/ovgs/ovg_nrw/j2016/14_A_3066_15_Beschluss_20160621.html

Schottmüller, H. (2008). ILIAS Formelfragen Erweiterung. Zugriff unter http://www.ilias.de/docu/goto_docu_file_1541_download.html

Ünlü, A. (2006). Estimation of careless error and lucky guess probabilities for dichotomous test items: A psychometric application of a biometric latent class model with random effects. *Journal of Mathematical Psychology*, *50*, 309–328. doi:10.1016/j.jmp.2005.10.002

VG Arnsberg, Urteil vom 17. April 2012. Aktenzeichen 9 K 399/11. Zugriff unter https://openjur.de/u/455037.print

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511618765

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201. doi:10.1111/j.1745-3984.1987.tb00274.x

Zentrum für Multimedia in der Lehre. (O. D.). *E-Assessment*. Universität Bremen. Zugriff unter http://www.eassessment.uni-bremen.de/fragetypen.php

Klausuren im sogenannten Antwort-Wahl-Format (AW-Klausuren), bei dem die richtige Antwort aus einer Reihe von vorgegebenen Alternativen auszuwählen ist, sind im Zeitalter der Digitalisierung zu einem festen Bestandteil vieler Prüfungsverfahren geworden. Problematisch an diesem Format ist allerdings die oft hohe Ratewahrscheinlichkeit.

Auf der Grundlage einer systematischen Analyse zeigen die Autoren, wie das Problem der Ratewahrscheinlichkeit begründbar, universell und gerichtsfest gelöst werden kann. Kernpunkt der Lösungsstrategie ist es, die Punktwerte für die verschiedenen Notenstufen an die Ratewahrscheinlichkeit der verwendeten Aufgaben anzupassen (ratekorrigierte Bestehens- und Notengrenzen).

Im theoretischen ersten Teil werden die kognitionspsychologischen und testtheoretischen Grundlagen von Aufgaben im AW-Verfahren skizziert und ein wahrscheinlichkeitstheoretisches Modell für das Antwortverhalten formuliert. Dieses Modell wird in Teil II auf die wichtigsten Aufgabenformate angewandt, die Berechnung der Ratekorrektur abgeleitet und nachvollziehbar dargestellt. Teil III enthält eine Kurzreferenz für die praktische Behandlung dieser Aufgabenformate.

Das Handbuch richtet sich dabei sowohl an Prüfende, als auch an alle mit Prüfungsangelegenheiten befassten Personen, wie Mitarbeitern in Prüfungsämtern und Prüfungsausschüssen, Juristen und nicht zuletzt: interessierte Prüflinge.

ISBN 978-3-86829-873-4