Electric Paper
EVALUATIONSSYSTEME

2017/03

Anleitung für die Gestaltung und Auswertung von Prüfungen mit EvaExam V7.1

Prof. Dr. Josef Lukas

Impressum

Electric Paper Evaluationssysteme GmbH

Konrad-Zuse-Allee 13
21337 Lüneburg
Deutschland

Telefon: +49 4131 7360 0
Telefax: +49 4131 7360 60
E-Mail: info@evasys.de

Geschäftsführer: Sven Meyer

USt-IdNr.: DE 179 384 158
Handelsregister: HRB-Nr. 1604, Lüneburg

Autor: Prof. Dr. Josef Lukas

Inhaltsverzeichnis

1. Vorbemerkung

Für das Erstellen, Durchführen und Auswerten von Prüfungen im Antwort-Wahl-Verfahren (*Multiple Choice-Prüfungen*) bietet EvaExam in der Version V7.1 vielfältige Möglichkeiten: Es gibt verschiedene Aufgabenformate und Aufgabentypen (Single Choice, Multiple Choice, Wahr/Falsch, Zuordnungsfragen, offene Fragen und einige andere), die Prüfungen können papierbasiert oder als Online-Prüfungen durchgeführt werden, die Vergabe von Punkten ist sehr flexibel und lässt viel Spielraum für die individuelle Gestaltung der Bewertung von Prüfungsleistungen. Die vorliegende Anleitung enthält Hinweise für eine sinnvolle und sachgerechte Nutzung der Möglichkeiten, die EvaExam bietet. Im Zentrum stehen dabei vor allem drei Fragen, die beim Erstellen von Klausuren im Antwort-Wahl-Verfahren zu beantworten sind:

- Wie sollen Punkte für richtige bzw. falsche Antworten vergeben werden? Kann, darf oder muss man Minuspunkte für Falschantworten vergeben?
- Wie berücksichtigt man am besten die Ratewahrscheinlichkeit bei Aufgaben im Antwort-Wahl-Verfahren?
- Wie berechnet man sinnvolle Bestehens- und Notengrenzen?

Der Aufbau dieser Anleitung ist so gestaltet, dass zunächst die wichtigsten Regeln aus der psychologischen Testtheorie für die Vergabe von Punkten (sog. *Scoring*-Verfahren) kurz charakterisiert werden. Für jedes in EvaExam verfügbare Aufgabenformat werden dann die jeweiligen Vor- und Nachteile der verschiedenen Scoring-Verfahren konkret erläutert. Der dritte Teil enthält schließlich Anleitungen für die Berechnung von Bestehens- und Notengrenzen sowie eine Erläuterung zum *Anpassungstool für Notenschlüssel,* einem Plug-in, das diese Berechnung in vielen Fällen automatisch vornehmen kann.

Um die hier vorliegende Anleitung kurz, übersichtlich und anwenderfreundlich gestalten zu können, haben wir auf eine Begründung für die vorgeschlagenen Auswertungsverfahren weitgehend verzichtet. Die theoretischen Grundlagen für diese Anleitung sowie ihre kognitionspsychologische, testtheoretische und statistische Begründung sind in dem Handbuch von Lukas, Melzer, Much & Eisentraut (2017) ausführlich dargelegt. Dieses Handbuch steht im Internetportal des Zentrums für Multimediales Lehren und Lernen (@LLZ) der Martin-Luther-Universität Halle-Wittenberg (www.llz.uni-halle.de) kostenlos zum Download zur Verfügung und wird allen Anwendern von EvaExam empfohlen, die sich eingehender über die hier besprochenen Methoden informieren wollen.

2. Scoring-Verfahren

Bei der Vergabe von Punkten für Aufgabenlösungen, dem sog. *scoring*, gibt es im Wesentlichen drei klassische Verfahren:

- das *Standardscoring*, bei dem es für jede richtige Antwort einen (oder auch mehrere) Punkte gibt und für falsche oder fehlende Antworten keine Punkte,
- das *formula scoring*, bei dem es für jede richtige Antwort einen (oder mehrere) Punkte gibt, für jede falsche Antwort aber Minuspunkte und keine Punkte für fehlende Antworten, sowie

- das *testlet scoring*, bei dem mehrere Einzelfragen zu einer Fragengruppe (dem sog. *testlet*) zusammengefasst werden und Punkte nur für die gesamte Gruppe vergeben werden.

2.1. Standardscoring

Das *Standardscoring* ist das einfachste Verfahren, es entspricht am ehesten der Intention, das Gesamtergebnis einer Prüfung durch die Anzahl der richtig beantworteten Fragen zu repräsentieren und es gilt allgemein als das Verfahren der Wahl. Wann immer es möglich ist, sollte das *Standardscoring* gewählt werden: Ein Punkt für jede richtige Antwort, kein Punkt für alle anderen Antworten (falsche Antworten oder fehlende Antworten). Statt einem Punkt können für eine richtige Antwort auch *mehrere* Punkte vergeben werden, die einzelnen Aufgaben (Items) können dadurch in ihrer Bedeutung für die Prüfung *gewichtet* werden. Die Frage, ob und falls ja in welcher Situation unterschiedliche Gewichte für die einzelnen Aufgaben sinnvoll sind und welche Auswirkungen die Gewichte auf das Gesamtergebnis haben, ist schwieriger zu beantworten, als es auf den ersten Blick erscheinen mag. Wer das genauer wissen möchte, kann sich an einem der Standardwerke zur Testtheorie orientieren (z.B. Lord & Novick, 1968). In Zweifelsfällen macht man nicht viel falsch, wenn man auf eine Gewichtung verzichtet und für jede richtige Antwort einen Punkt vergibt.

Alternativen zum *Standardscoring* werden vor allem mit dem Hinweis auf das Problem der Ratewahrscheinlichkeit motiviert. Als Ratewahrscheinlichkeit bezeichnet man die Wahrscheinlichkeit dafür, dass bei einer rein zufälligen Wahl der Antwort (z.B. durch Münzwurf oder Würfeln) die richtige Antwort getroffen wird. Da die Ratewahrscheinlichkeit gerade bei Aufgaben im Antwort-Wahl-Verfahren typischerweise hoch ist, können Punkte in nennenswertem Umfang auch durch blindes Raten erworben werden. *Formula scoring* und *testlet scoring* sollen dem entgegenwirken: Durch die Vergabe von Maluspunkten für Falschantworten beim *formula scoring* soll das Raten demotiviert (bzw. in seiner Auswirkung korrigiert) werden (siehe Abschnitt 2.2). Beim *testlet scoring* wird demgegenüber versucht, den Erwerb von Punkten durch Raten zu erschweren (siehe Abschnitt 2.3). Beide Strategien haben - wie in den beiden nächsten Abschnitten erläutert wird - gravierende Nachteile, sodass wir insgesamt empfehlen, am *Standardscoring* festzuhalten und das Problem der Ratewahrscheinlichkeit durch eine Anpassung der Bestehens- und Notengrenzen zu lösen (siehe Abschnitt 4). In EvaExam steht dafür das *Anpassungstool für Notenschlüssel* zur Verfügung, das alle Berechnungen automatisch vornimmt (siehe Abschnitt 4.2).

2.2. Formula scoring

Beim *formula scoring* (Holzinger, 1924) wird für jede richtige Antwort ein Punkt vergeben. Für fehlende Antworten gibt es keine Punkte, für falsche Antworten werden Punkte abgezogen (Minuspunkte, Maluspunkte) und zwar genau $\frac{g}{1-g}$ - viele Punkte. Dabei bezeichnet g die Ratewahrscheinlichkeit. Je größer die Ratewahrscheinlichkeit ist, desto mehr Punkte werden für „falsches Raten“ abgezogen.

Die Idee beim *formula scoring* ist eine Art „Ratekorrektur“: Wer die Antwort auf eine Frage nicht weiß, hat einen Erwartungswert von 0 Punkten, unabhängig davon, ob er die Frage nicht beantwortet oder es mit Raten versucht. Wer nicht antwortet, bekommt auf jeden Fall 0 Punkte. Wer rät, bekommt mit Wahrscheinlichkeit g einen Punkt und

mit Wahrscheinlichkeit $1 - g$ einen Punktabzug von $\frac{g}{1-g}$ - vielen Punkten, im statistischen Mittel also ebenfalls 0 Punkte. Diese – von der Idee her sehr intelligente – Art der Ratekorrektur hat allerdings (mindestens) drei große Nachteile:

- Maluspunkte bei Prüfungen sind juristisch zumindest umstritten. Es gibt eine ganze Reihe von Gerichtsurteilen, wonach Punkte, die durch richtige Antworten erworben wurden, nicht durch eine falsche Antwort wieder entzogen werden dürfen.
- Die Praxis von Maluspunkten richtet sich häufig nicht nach der oben angegebenen Formel für das *formula scoring,* sondern besteht oft im naiven Abzug von halben oder ganzen Punkten ohne Berücksichtigung der genauen Ratewahrscheinlichkeit.
- Und schließlich vergrößert das *formula scoring* den Fehler zweiter Art durch einen Punktabzug, der dadurch entsteht, dass eine „eigentlich" gut beherrschte Aufgabe durch einen „Schusselfehler" falsch beantwortet wurde.

Wer also trotz der juristischen Bedenken in einer Prüfung eine Ratekorrektur durch Maluspunkte für Falschantworten einsetzen möchte, sollte dies auf jeden Fall nach der Formel für das *formula scoring* tun. Im Abschnitt 3 erklären wir genauer, wie das in EvaExam geht. In den meisten Fällen dürfte es allerdings besser sein, auf Maluspunkte zu verzichten, das Standardscoring von Abschnitt 2.1 anzuwenden und anschließend die Bestehens- und Notengrenzen an die Ratewahrscheinlichkeit anzupassen.

2.3. Testlet scoring

Beim *testlet scoring* werden mehrere Aufgaben zu einer Fragengruppe (einem *testlet*) zusammengefasst. Die Punktvergabe erfolgt für die Fragengruppe insgesamt, z.B. so, dass Punkte erst dann vergeben werden, wenn *alle* Fragen der Gruppe richtig beantwortet wurden (siehe z.B. Millsap, 2011, S. 147).

Das *testlet scoring* ist außerordentlich effektiv, wenn es nur darum geht, die Wahrscheinlichkeit dafür zu senken, dass Punkte durch blindes Raten erworben werden. Der Fehler erster Art (die Ratewahrscheinlichkeit) lässt sich damit erfolgreich reduzieren - allerdings auf Kosten des Fehlers zweiter Art, also der Wahrscheinlichkeit, Punkte durch „Schusselfehler" zu verlieren. Diese Wahrscheinlichkeit ist beim testlet scoring drastisch erhöht (Lukas et al., 2017).

In EvaExam ist *testlet scoring* in den beiden Aufgabentypen „Multiple Choice-Fragen (nur korrekte Antwortkombinationen bewerten)" und „Kprim-Fragen" realisiert und fest vorgegeben. Zu den Vor- und Nachteilen dieser Aufgabentypen gibt es weitere Informationen in Abschnitt 3.

3. Aufgabenformate in EvaExam

3.1. Single Choice-Fragen

Bei Single Choice-Fragen ist genau eine der k-vielen Antwortoptionen richtig, alle anderen sind falsch. Beim Anlegen der Frage wird für jede Antwortoption ein Punktwert zwischen -10 und +10 eingegeben.

Standardscoring: Für die (einzige) richtige Antwortmöglichkeit wird der Punktwert 1 eingetragen (oder ein beliebiger positiver Wert > 0, wenn die Frage gewichtet werden soll), für alle falschen Antwortmöglichkeiten (die Distraktoren) wird der Punktwert 0 eingetragen.

Formula scoring: Für die richtige Antwortmöglichkeit wird der Punktwert 1 eingetragen, für alle falschen Antwortoptionen wird (wegen $g \coloneqq 1/k$) der Punktwert $-1/(k-1)$ eingetragen. Bei $k = 4$ Antwortoptionen beträgt der Punktwert für jede falsche Antwort also -1/3, bei $k = 5$ Antwortoptionen -1/4 usw. Wenn die Frage gewichtet werden soll, sind alle Punktwerte mit demselben Faktor zu multiplizieren. Bei $k = 3$ Antwortoptionen und dem Gewichtungsfaktor 5 als Beispiel sind als Punktwerte anzugeben: 5 für die richtige Antwort und jeweils -2.5 für die beiden falschen Antworten.

3.2. Wahr/Falsch-Fragen

Wahr/Falsch-Fragen sind Single Choice-Fragen mit genau zwei Antwortoptionen. Entsprechend ist nach Abschnitt 3.1 zu verfahren mit $k = 2$, das heißt:

Standardscoring: Für die richtige Antwortoption wird der Punktwert 1 eingetragen (oder ein beliebiger positiver Wert > 0, wenn die Frage gewichtet werden soll), für die falsche Antwortmöglichkeit wird der Punktwert 0 eingetragen.

Formula scoring: Für die richtige Antwortoption wird der Punktwert 1 eingetragen, für die falsche Antwortmöglichkeit der Punktwert -1. Soll die Frage gewichtet werden (mit dem Faktor m, wobei m größer als 0 sein muss und höchstens den Wert 10 haben darf), sind als Punktwerte m bzw. $-m$ anzugeben.

3.3. Multiple Choice-Fragen (einzelne Antwortoptionen bewerten)

Bei Multiple Choice-Fragen vom Typ „einzelne Antwortoptionen bewerten“ müssen Prüflinge für jede einzelne der vorgegebenen Antwortoptionen entscheiden, ob sie sie für „zutreffend“ halten und ankreuzen, oder ob sie nicht zutrifft und deshalb nicht angekreuzt werden soll. Wie viele der vorgegebenen Antwortoptionen richtig sind (zutreffen) und wie viele nicht zutreffen (Distraktoren) ist den Prüflingen im Allgemeinen nicht bekannt. Im Extremfall können alle Antwortoptionen zutreffen oder auch alle falsch sein.

Eine „Stimmenthaltung“ (durch Nichtbeantworten) ist bei diesem Fragetyp für die Prüflinge *nicht* möglich, da ein Nicht-Ankreuzen immer als Entscheidung für „trifft nicht zu“ gewertet wird (im Unterschied zu den Wahr/Falsch-Fragen, bei denen es jeweils drei Reaktionsmöglichkeiten gibt: Wahr – Falsch – keine Entscheidung). Als „richtig“ gilt eine Antwort dann, wenn entweder eine zutreffende Antwortoption angekreuzt wurde oder ein Distraktor nicht angekreuzt wurde (!). Entsprechend gilt eine Antwort als falsch, wenn entweder ein Distraktor angekreuzt oder eine korrekte Antwortoption nicht angekreuzt wurde. Die gelegentlich anzutreffende Praxis, Punkte nur für das Ankreuzen zu vergeben, ist ausdrücklich *nicht* empfehlenswert (für eine ausführliche Diskussion dazu siehe Lukas et al., 2017).

Standardscoring: Bei jeder richtigen Antwortoption wird für das Ankreuzen der Punktwert 1 eingetragen (oder ein beliebiger positiver Wert > 0, wenn die Frage gewichtet

werden soll), und für das Nicht-Ankreuzen der Punktwert 0. Bei den falschen Antwortoptionen ist es genau umgekehrt: Für das Ankreuzen gibt es 0 Punkte, für das Nicht-Ankreuzen einen (oder mehrere) Punkte. Von den beiden Punktwerten jeder Antwortoption ist also beim Standardscoring immer einer gleich 0 und der andere größer als 0.

Formula scoring: Für jede richtige Antwort wird der Punktwert 1 eingetragen, für jede falsche Antwort der Punktwert -1. Die beiden Werte jeder Antwortmöglichkeit lauten also (1; -1) für die zutreffenden Antwortmöglichkeiten und (-1; 1) für die Distraktoren. Soll die Frage gewichtet werden (mit dem Faktor m, wobei m größer als 0 sein muss und höchstens den Wert 10 haben darf), sind als Punktwerte m und $-m$ anzugeben.

Multiple Choice-Fragen vom Typ „einzelne Antwortoptionen bewerten“ sind wegen ihrer großen Flexibilität vielfältig einsetzbar und haben viele Vorteile:

- Sie eignen sich für alle Arten von Klassifikationsaufgaben, bei denen für eine Reihe von Aussagen oder Objekten entschieden werden soll, ob sie zur Kategorie A oder B gehören. Diese beiden Kategorien sind frei wählbar. Sehr häufig werden die Kategorien „wahr“ und „falsch“ verwendet, bei denen für jede Aussage entschieden werden soll, ob sie zutrifft oder nicht. Es sind aber auch alle anderen Arten von dichotomen Klassifikationsaufgaben realisierbar.
- Anders als bei Single Choice-Aufgaben muss nicht genau eine Aussage zur Kategorie A und alle anderen zur Kategorie B gehören. Zutreffende und nicht-zutreffende Aussagen können in beliebiger Anzahl und Kombination gemischt werden. Das erleichtert die Formulierung von sinnvollen Antwortoptionen erheblich.
- Während es sich bei Single Choice-Fragen um Diskriminationsaufgaben handelt, bei denen unter allen vorgegebenen Antwortoptionen lediglich diejenige herauszusuchen ist, die *am ehesten* für die Kategorie A in Frage kommt, muss bei Multiple Choice-Fragen vom Typ „einzelne Antwortoptionen bewerten“ für jede einzelne Antwortoption getrennt entschieden werden, ob sie zur Kategorie A oder zur Kategorie B gehört. Das liefert deutlich mehr Informationen über das Wissen des Prüflings.

Dass angesichts dieser offensichtlichen Vorteile Multiple Choice-Fragen vom Typ „einzelne Antwortoptionen bewerten“ in der Praxis häufig gemieden werden, liegt vermutlich an der hohen Ratewahrscheinlichkeit bei diesem Fragentyp. Für jede einzelne Entscheidung liegt die Wahrscheinlichkeit für eine richtige Antwort bei einer rein zufälligen Auswahl (z.B. Münzwurf) bereits bei ½. Wer ohne jedes Wissen einfach nur blind rät, erhält im statistischen Mittel 50% der maximal erreichbaren Punkte und damit die Bestehensgrenze für viele Klausuren im akademischen Bereich. Das ist natürlich nicht akzeptabel. Maßnahmen zur Ratekorrektur sind unerlässlich, wenn dieser Fragentyp verwendet werden soll. Man verwendet dafür entweder das oben erwähnte *formula scoring* oder – noch besser – man passt die Bestehens- und Notengrenzen an die hohe Ratewahrscheinlichkeit an (Abschnitt 4). Damit ist dieser Fragentyp uneingeschränkt in Prüfungen einsetzbar.

3.4. Multiple Choice-Fragen (nur korrekte Antwortkombination bewerten)

Es gibt in der Praxis noch einen anderen Ansatz, mit der hohen Ratewahrscheinlichkeit bei Multiple Choice-Fragen umzugehen: Man bewertet nicht jede einzelne Antwort zu jeder Antwortoption, sondern nur das Gesamtmuster aller Antworten bei einer Frage (*testlet scoring*). Punkte bekommt bei diesem Fragentyp nur, wer *alle* Antwortoptionen richtig beantwortet hat, das heißt: alle vom Prüfer als „richtig“ gekennzeichneten Alternativen angekreuzt hat, und *nur* diese. Ist auch nur ein Kreuz falsch gesetzt oder fehlt ein Kreuz, werden 0 Punkte vergeben. EvaExam läßt sogar die Vergabe von Minuspunkten (Strafpunkten) für fehlerhafte Antworten zu. Das ist aber bestenfalls in experimentellen Situationen sinnvoll. Bei realen Prüfungen sollte man davon keinesfalls Gebrauch machen. Dass ein Prüfling, der (fast) alles richtigmacht und lediglich bei einer Antwortoption falsch antwortet für die Gesamtaufgabe Minuspunkte erhält ist in keiner Situation vernünftig begründbar.

Das Problem der Ratewahrscheinlichkeit ist beim Fragetyp Multiple Choice-Fragen vom Typ „nur korrekte Antwortkombination bewerten“ effektiv gelöst. Bei vier Antwortoptionen beträgt die Wahrscheinlichkeit für die richtige Lösung der Gesamtaufgabe durch blindes Raten nur noch 1/16. Jede weitere Antwortoption halbiert die Ratewahrscheinlichkeit noch einmal. Der Preis dafür ist allerdings hoch:

- Es wird nicht unterschieden zwischen Prüflingen, die fast alles richtigmachen (bis auf einen Fehler) und Püflingen, die alles falsch machen. Beide erhalten 0 Punkte. Partielles Wissen wird nicht honoriert.
- Auch wer alles weiß und nur durch einen „Schusselfehler“ (*careless error*) eine der Antwortoptionen falsch beantwortet, bekommt 0 Punkte auf die Gesamtaufgabe. Im selben Maß, in dem der Fehler erster Art (Ratewahrscheinlichkeit) reduziert wird, erhöht sich dadurch der Fehler zweiter Art.

Dieser Fragentyp, bei dem das *testlet scoring* als Scoringverfahren fest integriert ist, sollte deshalb nur verwendet werden, wenn es dafür gute Gründe gibt, z.B., weil die einzelnen Antwortoptionen integraler Bestandteil des mit dieser Frage zu prüfenden Wissens sind. Wegen des Fehlers zweiter Art muss dabei die Anzahl der Antwortoptionen klein gehalten werden. Wenn es nur darum geht, die Ratewahrscheinlichkeit angemessen zu berücksichtigen, sind die hier besprochenen Methoden des *formula scorings* oder der Anpassung der Bestehens- und Notengrenzen beim Standardscoring im Allgemeinen besser geeignet als Methoden des *testlet scorings*.

3.5. Offene Fragen und segmentierte offene Fragen

Offene Fragen sind unter dem Gesichtspunkt der Scoring-Verfahren unkompliziert. Für jede Frage wird die maximale Punktzahl festgelegt. Für jede Antwort vergibt der Prüfer je nach Güte der Antwort die volle Punktzahl oder Teilpunkte. Die Ratewahrscheinlichkeit wird in der Regel bei offenen Fragen mit 0 angenommen.

3.6. Kprim-Fragen

Kprim-Fragen sind im Prinzip Multiple Choice-Fragen vom Typ „nur korrekte Antwortkombinationen bewerten“ mit zwei Besonderheiten:

- Es gibt immer genau vier Antwortalternativen.
- Sind alle vier Antwortalternativen richtig beantwortet, wird die volle Punktzahl vergeben. Bei einem Fehler (drei richtigen) wird die halbe Punktzahl vergeben. Bei mehr als einem Fehler werden keine Punkte vergeben.

Bei Kprim-Fragen ist damit ebenfalls ein *testlet scoring* fest vorgegeben. Dieses spezielle Format wird vor allem in der Medizin verwendet und soll die in Abschnitt 3.4 erwähnten Nachteile der Multiple Choice-Fragen vom Typ „nur korrekte Antwortkombinationen bewerten“ etwas abmildern. In der Praxis hat sich das nach Ansicht der Befürworter von Kprim-Fragen bewährt, aus testtheoretischer Perspektive sind Kprim-Fragen aber eher ein etwas provisorischer „work around“ als eine systematische Lösung für den Umgang mit Ratewahrscheinlichkeiten.

3.7. Zuordnungsfragen

Zuordnungsfragen haben eher spielerischen Charakter. Für Prüfungen sind sie nur eingeschränkt geeignet, weil in der Regel nicht ganz klar ist, welches Wissen damit abgefragt wird. Das größte formale Problem besteht darin, dass die Antwortmöglichkeiten stark und systematisch voneinander abhängig sind und beim Beantworten eine getroffene Wahl alle weiteren beeinflusst. Wer z.B. bei der ersten Zuordnung einen Fehler macht, kann nicht mehr alle anderen richtigmachen. Bei der letzten Zuordnung gibt es ohnehin keine Wahlmöglichkeit mehr usw. Da jede Zuordnungsfrage viel präziser durch Single Choice- oder Multiple Choice-Fragen realisiert werden kann, ist dieser Fragentyp auch ohne weiteres verzichtbar.

4. Bestehens- und Notengrenzen

4.1. Anpassung der Bestehens- und Notengrenzen

Für das Bestehen einer Prüfung wird in der Regel eine Mindestanzahl von Punkten gefordert, typischerweise als Prozentsatz der maximal erreichbaren Punkte. Ein häufiger Wert dabei ist eine Bestehensgrenze von 50%. In analoger Weise können Grenzwerte für Notenstufen definiert werden, z.B. 95% der Gesamtpunktzahl für die Note 1.0. In EvaExam steht dafür eine eigene Notenschlüsselverwaltung zur Verfügung, mit der jeder Prüfung ein eigener Notenschlüssel zugeordnet werden kann.

Bei Aufgaben mit einer nennenswerten Ratewahrscheinlichkeit (z.B. bei praktisch allen Fragen im Antwort-Wahl-Verfahren) macht es jedoch wenig Sinn, als Notengrenze einen festen Prozentsatz der Maximalpunktzahl anzugeben. Es hängt ja stark von der Ratewahrscheinlichkeit ab, wie viel man tatsächlich wissen muss, um „50% der maximalen Punktzahl“ zu erreichen. Bei offenen Fragen (mit einer verschwindenden Ratewahrscheinlichkeit) muss man dafür vermutlich etwa die Hälfte des Stoffes beherrschen. Bei Wahr/Falsch-Fragen dagegen erreicht man eine Quote von 50% im Durchschnitt aber bereits ohne jedes Wissen nur durch blindes Raten.

Die Lösung für dieses offenkundige Problem besteht darin, die Bestehens- und Notengrenzen so zu verändern, dass die Ratewahrscheinlichkeit mitberücksichtigt wird

(„Ratekorrektur“). Wir ersetzen dabei das Prinzip „bestanden hat, wer 50% der maximalen Punktzahl erreicht“ durch die Regel: „bestanden hat, wer 50% *weiss*“. Da das Wissen aber nicht direkt beobachtbar ist, operationalisieren wir „50% Wissen“ durch den „Erwartungswert der Punkte in der Klausur für jemanden, der 50% weiss“.

Ein konkretes Beispiel aus der Praxis soll das erläutern: Klausuren im Medizinischen Staatsexamen verwenden fast ausschließlich Single Choice-Fragen mit jeweils fünf Antwortalternativen. Die Ratewahrscheinlichkeit wird deshalb für jede Frage mit 0.2 angenommen. Da das Standardscoring verwendet wird, beträgt der Erwartungswert für Prüflinge, die genau die Hälfte des Stoffes sicher beherrschen, 60% der maximalen Punktzahl (die Hälfte der maximalen Punktzahl erreichen sie durch ihr Wissen, bei der anderen Hälfte raten sie und erreichen dabei durchschnittlich 1/5, also weitere 10% der Punkte). Für das Bestehen der Prüfung sind deshalb nicht 50%, sondern 60% der erreichbaren Punkte erforderlich.

Der Vorteil dieses Verfahrens besteht vor allem darin, dass man z.B. in Prüfungsordnungen feste Notenschlüssel für das Bestehen und die verschiedenen Notenstufen angeben kann, die aber für jede Klausur an die Ratewahrscheinlichkeit der verwendeten Aufgaben angepasst werden. Je höher die Ratewahrscheinlichkeit, desto höher werden die Bestehens- und Notengrenzen. Anweisungen zur konkreten Berechnung der Ratekorrektur sind in Lukas et al. (2017) enthalten.

4.2. Das *Anpassungstool für Notenschlüssel* als Plug-in zu EvaExam

In EvaExam steht für die Berechnung der ratekorrigierten Bestehens- und Notengrenzen ein „*Anpassungstool für Notenschlüssel*“ als Plug-in zur Verfügung. Ausgangspunkt ist ein (allgemeiner) Notenschlüssel, der in der Notenschlüsselverwaltung eingerichtet wird und der die Notenstufen und die dazugehörigen Punktintervalle definiert. Das *Anpassungstool für Notenschlüssel* berechnet daraus für jede Klausur die ratekorrigierten neuen Grenzen und die daraus resultierenden Noten für alle Teilnehmer. Die Klausur kann dabei (fast) beliebig zusammengesetzt sein aus unterschiedlichen Fragentypen mit unterschiedlicher Ratewahrscheinlichkeit. Vorausgesetzt wird lediglich, dass bei allen Fragen das Standardscoring angewendet wird. Das ist allerdings keine echte Einschränkung, da Anwender, die *formula scoring* oder *testlet scoring* verwenden, sich für eine andere Methode der Ratekorrektur entschieden haben, so dass eine Anpassung der Bestehens- und Notengrenzen nicht nötig ist. Die Anwendung des *Anpassungstools für Notenschlüssel* ist deshalb nur sinnvoll, wenn die Klausur keine Multiple Choice-Fragen mit *testlet scoring*, keine Kprim-Fragen und keine Zuordnungsfragen enthält. Single Choice-, Multiple Choice (mit Bewertung der einzelnen Antworten), offene Fragen und segmentierte offene Fragen können dagegen beliebig gemischt (und auch unterschiedlich gewichtet) werden.

Ein zusätzlicher Vorteil für die Methode „Standardscoring mit Ratekorrektur der Bestehens- und Notengrenzen“ ist, dass zusätzlich ein Wert für den Fehler zweiter Art (Wahrscheinlichkeit für eine falsche Antwort trotz sicheren Wissens) angegeben werden kann. Es wird empfohlen, hier einen Wert zwischen 0.01 und 0.05 anzugeben. Da bei der Ratekorrektur das „Raten“ explizit berücksichtigt wird, spricht vieles dafür, aus Gründen der Fairness auch die Möglichkeit von „Schusselfehlern“ in die korrigierten Bestehens- und Notengrenzen mit einzubeziehen.

5. Literatur

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

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

Lukas, J., Melzer, A., Much, S. & Eisentraut, S. (2017). *Auswertung von Klausuren im Antwort-Wahl-Format*. Zentrum für Multimediales Lehren und Lernen (@LLZ), Martin-Luther-Universität Halle-Wittenberg.

Millsap, R.E. (2011). Statistical approaches to measurement invariance. New York: Routledge.