universität wien

Diplomarbeit

Titel der Arbeit

Multiple-Evaluation

Pilotstudie zu einem neuen Antwortformat

Verfasserin

Livia Maria Jenner

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im Jänner 2012

Studienkennzahl:	298
Studienrichtung:	Psychologie
Betreuer:	Univ.-Prof. Dr. Mag. Klaus D. Kubinger

Danksagung

An dieser Stelle möchte ich allen Personen danken, die mich während meiner Studienzeit und beim Schreiben meiner Diplomarbeit unterstützt haben.

An erster Stelle danke ich Herrn Univ.-Prof. Dr. Mag. Klaus Kubinger für seine fachliche Betreuung sowie Herrn Prof. Dr. Jochen Musch und Herrn Prof. Dr. Erich Neuwirth für ihre Anregungen.

Mein besonderer Dank gilt der Direktion und den Lehrerinnen des Gymnasiums, die diese Testung erst ermöglicht haben sowie allen Schüler/innen, die mit Interesse und Elan bei der Testung dabei waren.

Außerdem möchte ich meiner Kollegin und Freundin Anita danken, nicht nur dafür, dass sie mich bei der Durchführung und dem Schreiben meiner Diplomarbeit unterstützt und motiviert hat, sondern auch dafür, dass sie während des gesamten Studiums immer meine erste Ansprechpartnerin und Vertraute war.

Kathi, Marianne, Nina, Paulina, Sabine und Steffi danke ich für ihre Freundschaft und, dass sie mir oftmals so geduldig zugehört haben sowie für die vielen schönen Stunden, die ich während meiner Studienzeit mit ihnen verbringen durfte.

Meinem Freund Elton danke ich dafür, dass er immer für mich da ist und mich vor allem auch in schwierigen Zeiten mit viel Geduld ermutigt hat weiter zu machen.

Nicht zuletzt möchte ich meinen Eltern danken, die mir nicht nur das Studium ermöglicht haben, sondern mich bis zum Abschluss meiner Diplomarbeit, wann immer sie konnten, unterstützt haben. Ihnen beiden soll meine Arbeit gewidmet sein.

Abstract

Multiple-Evaluation provides, unlike Multiple-Choice, a possibility to measure partial knowledge, by giving the person 100 percent, which can be distributed on the given answers. It is based on a logarithmic scoring rule and can lead to a negative item score. Through wild guessing there is a risk of getting penalty-points. If one's knowledge is estimated and reproduced realistically, the score can be maximized. Through this a reduction of the guessing effect and an enhancement of the reliability and the validity of a test are expected. However, some authors postulate personality- and gender-influences on the way a person answers in the Multiple-Evaluation response-format. This study explores the influence of personality and gender on the answering behavior in a multiple-evaluation testing situation. Therefore an achievement test and a personality questionnaire were applied to 162 pupils. According to Hansen (1971) the average item-certainty was calculated to measure the certainty someone displays in their response. Further the degree in which a person profits from a logarithmic scoring was computed. No meaningful correlations between the personality variables and the answering certainty or the profit from the logarithmic scoring rule were found. Concerning the sex differences, men are significantly more certain in their answers than women and women profit significantly more from a logarithmic scoring rule, which indicates that they estimate and reproduce their knowledge more realistically.

Abstract

Das Multiple-Evaluations Format bietet im Gegensatz zum Multiple-Choice Format die Möglichkeit partielles Wissen zu erfassen, indem der Testperson 100 Prozent zu Verfügung gestellt werden, die sie auf die ihr vorgegebenen Antwortmöglichkeiten verteilen kann. Das Antwortformat basiert auf einer logarithmischen Auswertung, bei der es auch zu negativen Itemscores kommen kann. Durch „wildes“ Raten geht die Testperson ein Risiko ein Strafpunkte zu erhalten. Gibt sie ihr Wissen korrekt und unverfälscht wieder, kann sie ihren Score maximieren. Durch das spezielle Responseformat und die Art des Scorings werden eine Reduktion des Rateeffekts sowie eine Erhöhung der Reliabilität und der Validität erwartet. Einige Autoren postulieren jedoch Persönlichkeitseinflüsse auf das Antwortverhalten im Multiplen-Evaluations Format. Diese Arbeit untersucht den Einfluss von Geschlecht und Persönlichkeit auf das Antwortverhalten bei Vorgabe eines Tests im Multiplen-Evaluations Format. Hierfür wurden ein Leistungstest im Multiplen-Evaluations Format sowie ein Persönlichkeitsfragebogen 162 Schüler/innen vorgegeben. Als Maß für die Antwortsicherheit einer Person wurde die durchschnittliche Itemsicherheit nach Hansen (1971) berechnet. Ferner wurde das Ausmaß des Profits, den die Testpersonen von einer logarithmischen Auswertung haben, ermittelt. Es konnten keine bedeutsamen Zusammenhänge zwischen den erhobenen Persönlichkeitsvariablen und der Antwortsicherheit beziehungsweise dem Ausmaß des Profits von einer logarithmischen Auswertung gefunden werden. Bezüglich der Geschlechtsunterschiede zeigte sich, dass Männer signifikant antwortsicherer sind als Frauen und Frauen signifikant mehr von einer logarithmischen Auswertung mit Strafpunkten profitieren, was darauf schließen lässt, dass sie ihr Wissen realistischer einschätzen und wiedergeben.

Inhalt

I. Einleitung 1

II. Theoretischer Teil 5

1. Möglichkeiten und Grenzen des Multiple-Choice Antwortformats 7

1.1 Strategien gegen den Rateeffekt 9

1.2 Partielles Wissen 13

2. Das Multiple-Evaluations Format 15

2.1 Probabilistisches Messen 15

2.2 Das Multiple-Evaluations Format als Spezialfall des probabilistischen Messens 16

2.3 Logarithmische Verrechnung 17

2.4 Vor- und Nachteile des Multiplen-Evaluations Formats 19

3. Einflussfaktoren auf das Antwortverhalten beim probabilistischen Messen 21

3.1 Antwortverhalten beim probabilistischen Messen 21

3.2 Situative Einflüsse auf das Antwortverhalten 22

3.3 Persönlichkeitseinflüsse auf das Antwortverhalten 22

3.4 Einflüsse des Geschlechts auf das Antwortverhalten 23

III. Empirischer Teil 25

4. Zielsetzung der Untersuchung 27

4.1 Hypothesen 27

4.1.1 Hypothesen zu Persönlichkeitseinflüssen auf die Antwortsicherheit 27

4.1.2 Hypothesen zu Geschlechtsunterschieden bezüglich der Antwortsicherheit 28

4.1.3 Hypothesen zu Persönlichkeitseinflüssen auf das Ausmaß des Profites durch eine logarithmische Auswertung 28

4.1.4 Hypothesen zu Geschlechtsunterschieden bezüglich des Ausmaßes des Profits von einer logarithmische Auswertung 28

4.1.5 Hypothesen zum Einfluss der Antwortsicherheit einer Person auf das Ausmaß ihres Profits durch eine logarithmische Auswertung 28

5. Methode 29

5.1 Untersuchungsplan 29
5.2 Erhebungsinstrumente 30
5.2.1 Satzergänzung (I-S-T 2000R) 30
5.2.2 HPI 31
5.3 Durchführung der Untersuchung 32
5.3.1 Instruktion und Übungsbeispiele 32
5.3.2 Testung 33
5.4 Stichprobe 33
6. Ergebnisse 35
6.1 Datenerfassung und Berechnung der Variablen 35
6.2 Hypothesenprüfung 35
6.2.1 Ergebnisse zum Einfluss der Persönlichkeitsvariablen auf die durchschnittliche Itemsicherheit nach Hansen (C_T) 35
6.2.2 Ergebnisse zum Geschlechtsunterschied bezüglich der durchschnittlichen Itemsicherheit 36
6.2.3 Ergebnisse zum Einfluss von Persönlichkeitsvariablen auf das Ausmaß des Profits von einer logarithmischen Auswertung 37
6.2.4 Ergebnisse zum Geschlechtsunterschied bezüglich des Ausmaßes des Profits von einer logarithmischen Auswertung 38
6.2.5 Ergebnisse zum Einfluss der durchschnittlichen Itemsicherheit (C_T) auf das Ausmaß des Profits von einer logarithmischen Auswertung 38
7. Diskussion der Ergebnisse und Ausblick 40
7.1 Diskussion der Ergebnisse 40
7.1.1 Diskussion der Ergebnisse zu Persönlichkeitseinflüssen auf das Antwortverhalten 40
7.1.2 Diskussion der Ergebnisse zum Einfluss des Geschlechts auf das Antwortverhalten 41
7.1.3 Diskussion der Ergebnisse zum Einfluss der Antwortsicherheit auf das Ausmaß des Profits von einer logarithmischen Auswertung 42

7.2 Ausblick 43
8. Zusammenfassung 45
9. Literaturverzeichnis V
10. Tabellenverzeichnis IX
11. Anhang XI
11.1 Instruktion XI
11.2 Instruktionsblatt XXI
Lebenslauf XXIII

I. Einleitung

In der Psychologischen Diagnostik spielt die Wahl des Antwortformats bei der Itemgestaltung eine entscheidende Rolle, da es bestimmt, in welcher Art und Weise die von der Testperson gesetzten Reaktionen auf ein Item registriert werden (Seiwald, 2003).

Das Multiple-Choice Antwortformat dominiert vor allem auf Grund seiner ökonomischen Durchführung und Auswertung das Feld der modernen Testkonstruktion, es bringt jedoch auch Nachteile mit sich. So ist es in diesem Format unter anderem nicht möglich, partielles Wissen zu erfassen, da sich die Testpersonen, auch wenn sie über Teilwissen verfügen, für eine Antwortmöglichkeit entscheiden müssen.

Probabilistisches Messen bietet die Möglichkeit, bei Vorgabe eines gebundenen Antwortformats partielles Wissen zu erfassen. Die Testperson hat 100 Prozent zu Verfügung, die sie auf die Antwortmöglichkeiten verteilen kann. Die vergebenen Prozente sollen dabei anzeigen, wie wahrscheinlich die Testperson die jeweilige Antwortmöglichkeit für richtig hält.

Die vorliegende Arbeit beschäftigt sich mit dem Multiplen-Evaluations Format, welches einen Spezialfall des probabilistischen Messens darstellt und auf einer logarithmischen Auswertung basiert, bei der es zu negativen Itemscores kommen kann. Rät eine Person „wild“, geht sie das Risiko ein, auf Grund dieser Verrechnung Strafpunkte zu erhalten. Gibt sie ihr Wissen korrekt und unverfälscht wieder, kann sie ihren Score maximieren. Durch das spezielle Responseformat sowie durch die Art des Scorings werden eine Reduktion des Rateeffekts sowie eine Erhöhung der Reliabilität und der Validität eines Tests erwartet (Dirkzwager, 1996, 2003; Holmes, 2002).

Bezüglich der Validität eines Tests im Multiplen-Evaluations Formats konnten einige Autoren Ergebnisse finden, die darauf schließen lassen, dass neben dem Wissen einer Person auch Persönlichkeitsmerkmale einen Einfluss darauf haben, wie eine Person die Prozente auf die ihr vorgegeben Antwortmöglichkeiten verteilt (Hanse, 1971; Koehler, 1974; Lundeberg, Fox & Puncochar, 1994; Schaefer, Williams, Goodie & Campebell, 2004).

Im Rahmen dieser Arbeit werden mögliche Geschlechts- und Persönlichkeitseinflüsse auf das Antwortverhalten einer Person bei Vorgabe eines Tests im Multiplen-Evaluations Format überprüft. Theoriegeleitet wird das Antwortverhalten durch die Antwortsicherheit (durchschnittliche Itemsicherheit nach Hansen, 1971) der Testpersonen beschrieben. Außerdem wird überprüft, inwiefern Persönlichkeit und Geschlecht einen Einfluss darauf

haben, in welchem Ausmaß Testpersonen von einer logarithmischen Auswertung mit Strafpunkten profitieren bzw. ihnen Nachteile daraus erwachsen.

Im theoretischen Teil wird erläutert, welche Möglichkeiten und Grenzen das Multiple-Choice Antwortformat mit sich bringt und weshalb probabilistisches Messen für die Psychologische Diagnostik von Interesse ist. Es folgt eine detaillierte Beschreibung des Multiplen-Evaluations Antwortformats sowie der dazugehörigen logarithmischen Auswertung. Abschließend werden Möglichkeiten vorgestellt, das Antwortverhalten einer Testperson im Multiplen-Evaluations Format zu erfassen und mögliche Einflüsse des Geschlechts und der Persönlichkeit auf das Antwortverhalten diskutiert.

Im empirischen Teil der Arbeit werden die Hypothesen zu Persönlichkeits- und Geschlechtseinflüssen auf das Antwortverhalten definiert sowie die Durchführung der Testung im Detail beschrieben. Dem folgt eine ausführliche Darstellung der Ergebnisse. Abschließend werden die Ergebnisse interpretiert und diskutiert sowie ein Ausblick für weitere Studien gegeben.

II. Theoretischer Teil

1. Möglichkeiten und Grenzen des Multiple-Choice Antwortformats

Bei der Konstruktion und der Anwendung psychologisch-diagnostischer Verfahren muss der Itemgestaltung eine große Rolle beigemessen werden (Seiwald, 2003, S.23).

„*Vor allem das Antwortformat, also die Art und Weise, in der die von der Testperson gesetzten Reaktionen auf die Items registriert werden, bestimmt grundsätzliche Eigenschaften des Verfahrens.“* (Seiwald, 2003, S.23-24).

Nach Seiwald (2003) kann zwischen freiem und gebundenem Antwortformat unterschieden werden. Das freie Antwortformat zeichnet sich dadurch aus, dass die Testperson die Antwort auf das Item selbst formulieren muss. Bei gebundenen Antwortformaten wird pro Item eine bestimmte Anzahl an vorformulierten Antwortmöglichkeiten angeboten. Beim Multiple-Choice (MC) Format muss die Testperson im Fall von Leistungstests die vermeintlich richtige bzw. bei Persönlichkeitsfragebögen die passende, der Testperson am ehesten entsprechende, Antwortmöglichkeit wählen (vgl. Seiwald, 2003; Kubinger, 2009).

Tests im Multiple-Choice Format dominieren das Feld der modernen Testkonstruktion. Neben standardisierten Tests werden auch ad hoc Tests (*classroomtest, teachermadetests*), meist Prüfungsaufgaben, immer öfter im Multiple-Choce Format gestaltet. Die ökonomische Durchführung und Auswertung von Multiple-Choice Tests kann als einer der Hauptgründe für deren Beliebtheit gesehen werden (vgl. Lienert & Ratz, 1998; Schaefer, 1976). In vielen Fällen ist die Auswertung bereits maschinell, mittels „Einlesen“ durch den Computer, möglich (Seiwald, 2003). Die in der Anwendung ebenfalls ökonomischeren Gruppenverfahren sind meist im Multiple-Choice Format gestaltet und auch die immer häufiger eingesetzte Computerdiagnostik profitiert von diesem Antwortformat (Kubinger, 2009).

Außerdem kann beim Multiple-Choice Format eine höhere Verrechnungssicherheit als beim offenen Antwortformat erwartet werden. Da bei der Auswertung von Multiple-Choice Items keinerlei Ermessensfreiheit besteht, könnten Auswerter/innen nur auf Grund von Auswertungsfehlern zu unterschiedlichen Ergebnissen kommen. Somit ist das Gütekriterium der Objektivität bei dieser Art der Testvorgabe in hohem Ausmaß erfüllt (Kubinger, 2009, S.45).

Das Format bringt jedoch auch Nachteile mit sich. Im Gegensatz zu Fragen im offenen Antwortformat kann die Vorgabe von Antwortmöglichkeiten und der dadurch eingeschränkte „Handlungsspielraum“ der Testpersonen zu einem geringeren diagnostischen Informationsgehalt führen (Seiwald, 2003, S. 25).

Aus Experimenten der Allgemeinen Psychologie ist bekannt, dass das Wiedererkennen von gelernten Inhalten deutlich leichter fällt, als das Reproduzieren derselben (Zimbardo & Gerrig, 2008). Nach Kubinger (2009) ist das Multiple-Choice Antwortformat deshalb dann ungeeignet, wenn es um die freie und selbständige Wissenswiedergabe durch die Testperson geht.

Darüber hinaus wird die diagnostische Tragweite des Rateeffekts, welcher bei Leistungstests im Multiple-Choice Format erwartet werden muss, sehr oft unterschätzt. Unter Rateeffekt wird im Allgemeinen die a-priori Wahrscheinlichkeit verstanden, ein Item zufällig zu lösen. Auch wenn eine Person absolut kein Wissen bzw. keine Fähigkeit besitzt, so besteht immer eine Wahrscheinlichkeit von 1/k (k=Anzahl der Antwortmöglichkeiten), dass sie die richtige Antwortmöglichkeit durch zufälliges Raten wählt und das Item als „gelöst“ verrechnet wird. Die gängigsten psychologisch-diagnostischen Verfahren im Multiple-Choice Format haben fünf Antwortmöglichkeiten pro Item, eine richtige Antwortmöglichkeit sowie vier Distraktoren. Bei einem solchen Test beträgt die a-priori Wahrscheinlichkeit für ein Item die richtige Antwort zu erraten 1/5, also 20 Prozent. Somit kann eine Person ohne jegliche Fähigkeit im Durchschnitt jedes fünfte Item lösen. Das Problem wird dadurch verschärft, dass eine Testperson mit einem mittleren Fähigkeitsniveau meistens einzelne Distraktoren ausschließen kann. Dadurch wird die faktische Ratewahrscheinlichkeit höher als 20 Prozent. Ob eine Testperson mit moderaten Fähigkeiten einen Test besteht oder nicht, hängt davon ab, ob sie beim Raten Glück oder Pech hat (Kubinger, 2009).

Beim Multiple-Choice Format besteht also die Chance, nicht nur durch Fähigkeit bzw. Wissen, sondern auch durch Raten zu einer Lösung zu kommen, wobei dieser Aspekt nicht in die Auswertung eingeht. Es ist fraglich, inwiefern eine solche Art der Testung den zu testenden Personen zumutbar ist. Dadurch, dass einige Testpersonen es grundsätzlich ablehnen zu raten und andere nicht, kann das Kriterium der Testfairness als gefährdet angesehen werden (Kubinger, 2009).

1.1 Strategien gegen den Rateeffekt

Im Laufe der Jahre wurden einige Strategien entwickelt, um dem Rateeffekt bei Vorgabe des Multiple-Choice Formats entgegenzuwirken (vgl. Kubinger, 2009).

Da ein gegenläufiger Zusammenhang zwischen der Höhe des Rateeffekts und der Anzahl der Antwortmöglichkeiten besteht, scheint deren Erhöhung die einfachste mögliche Vorgehensweise, um den Rateeffekt zu reduzieren. Eine beliebige Erhöhung der Anzahl der Antwortmöglichkeiten ist jedoch nicht zielführend, da hierbei andere Phänomene wie Einflüsse der Merkfähigkeit, der Konzentration und der Leistungsmotivation zum Tragen kommen können (Kubinger, 2009). Kubinger (2009) weist außerdem darauf hin, dass eine übermäßige Erhöhung der Anzahl der Antwortmöglichkeiten die Testökonomie dahingehend gefährdet, dass der zeitliche Aufwand der Testbearbeitung übermäßig steigt. Des Weiteren stellt die Erstellung von „guten“ Distraktoren eine keineswegs triviale Aufgabe dar. Lienert und Raatz (1998) empfehlen, als Distraktoren häufig gegebene Antworten bei Ergänzungsaufgaben heran zu ziehen. Ergänzungsaufgaben sind solche, bei denen die Testperson eine Aufgabe durch ein Wort (Schlüsselwort) oder eine kurze Darstellung (zum Beispiel ein Symbol) beantwortet. Moreno, Martinez und Muniz (2006) konnten in einer Metaanalyse zeigen, dass fast alle Untersuchungen Ergebnisse liefern, die dafür sprechen, dass in den meisten Disziplinen drei Antwortmöglichkeiten pro Item ausreichen. Sie führen dies darauf zurück, dass die Konstruktion von mehreren, der Testperson plausibel und attraktiv erscheinenden, Distraktoren, oft nicht gelingt. In diesem Fall können die entwickelten Distraktoren von den Testpersonen, auch ohne dass sie die richtige Antwortmöglichkeit kennen, von vornherein ausgeschlossen werden. Die Autoren schlagen deshalb vor, so viele Distraktoren zu konstruieren, wie es der entsprechende Bereich zulässt, um die Item-Konstruktion nicht unnötig zu erschweren.

Eine Möglichkeit den Rateeffekt a-priori zu minimieren, ohne die Anzahl der Antwortmöglichkeiten zu erhöhen, ist die Verwendung von Items, bei denen zwei Antwortmöglichkeiten richtig sind. Das Item wird nur dann als gelöst verrechnet, wenn die Testperson beide richtigen Antwortmöglichkeiten und keine falsche Antwortmöglichkeit markiert hat. Eine Möglichkeit den Rateeffekt noch weiter herab zu setzten ist die Vorgabe von Items, bei denen beliebig viele Antwortmöglichkeiten richtig oder falsch sein können. Bei diesem Antwortformat, das bei Vorgabe von 5 Antwortmöglichkeiten kurz als „x aus 5“–Format bezeichnet wird, können auch gar keine oder alle Antwortmöglichkeiten richtig sein.

Eine Aufgabe wird dann als gelöst verrechnet, wenn von der Testperson alle richtigen, aber keine falschen Antwortmöglichkeiten gewählt wurden (Kubinger, 2009).

Ein weiterer Ansatz die Ratewahrscheinlichkeit a-priori zu verringern, ist der Versuch, die Testpersonen durch besondere Instruktionen oder Antwortmöglichkeiten, wie „ich weiß die Lösung nicht“ oder „keine Antwortmöglichkeit ist richtig“, vom Raten abzuhalten.(Kubinger, 2009). Wenn psychologisch-diagnostische Verfahren Items mit neutralen Kategorien, wie zum Beispiel „weiß nicht“, einsetzen, sind diese nach Bortz und Döring (2009) dann schwer auszuwerten, wenn viele Testpersonen diese Kategorie wählen und dies eventuell aus verschiedenen Motiven. Es können hierbei verschiedene Gründe unterschieden werden. So kann die Testperson etwas wirklich nicht wissen, sich unsicher sein, die Frage nicht beantworten wollen, oder zwischen mehreren Antwortmöglichkeiten schwanken. Die Autoren empfehlen deshalb eine solche Kategorie zu vermeiden oder, wenn ihre Verwendung unumgänglich ist, sie in der Instruktion zumindest so zu präzisieren, dass deutlich wird, was die Wahl der neutralen Kategorie ausdrückt. Betreffend der „Keine Antwortmöglichkeit ist die richtige“-Option, scheint es fraglich, inwiefern es ethisch vertretbar ist, falls tatsächlich für alle Aufgaben in den Antwortmöglichkeiten eine Lösung enthalten ist, den Testpersonen zu suggerieren, dass möglicherweise keine Antwortmöglichkeit stimmt (Kubinger, 2009). Kubinger (2009) stellt ferner die Frage, ob die Wahl solcher Antwortmöglichkeiten auch von Persönlichkeitsfaktoren abhängt. Da nicht feststellbar ist, ob alle Testpersonen in gleichem Maße darauf vertrauen, dass es honoriert wird, wenn sie ihr Nichtwissen zugeben, anstatt eine falsche Antwort infolge von Raten zu riskieren, bleibt diese Vorgehensweise mit einer gewissen Unsicherheit behaftet (Kubinger, 2009).

Das sequentielle Testen stellt einen neuen Ansatz dar, um die Ratewahrscheinlichkeit a-priori zu verringern. Der Testperson wird bei dieser Art des Testens jede Antwortmöglichkeit einzeln nacheinander dargeboten. Sie hat der Reihe nach pro Antwortmöglichkeit zu entscheiden, ob diese richtig oder falsch ist. Das Korrigieren bereits als falsch beurteilter Antwortmöglichkeiten ist dabei nicht möglich. Es stellt sich in diesem Zusammenhang die Frage, inwiefern es bei Tests, bei denen es um das Finden von „relativ bestpassenden“ Antwortmöglichkeiten geht, bei Vorgabe eines solchen sequentiellen Antwortformats nicht auch testpersonenabhängig sein kann, ob diese auf eine (noch) bessere Antwortmöglichkeit warten und dabei eventuell die vom Testautor vorgesehene Lösung verpassen (Kubinger, 2009).

Ratekorrekturen stellen eine Möglichkeit dar, die Ratewahrscheinlichkeit nachträglich zu berücksichtigen. Ziel der Ratekorrektur ist es, den Testrohwert auf den „wahren“ Wert zu reduzieren. Dies geschieht durch das nachträgliche Abziehen von Punkten, die auf Grund von zufällig richtigem Raten erzielt werden (Schaefer, 1976). Bei der einfachsten Form der nachträglichen Ratekorrektur bekommt die Testperson einen Punkt für jedes gelöste Item und 1/(k-1) Minuspunkte für jedes falsch beantwortete Item. Dieser Ansatz geht davon aus, dass alle falsch beantworteten Items das Produkt von zufälligem Raten sind und dass somit der entsprechende Anteil aller richtigen Antworten auch das Resultat von zufälligem Raten sein muss. Bei k=4 zum Beispiel, heben 3 falsche Antworten eine richtige auf. Meistens wird in diesem Fall den Personen die Möglichkeit gegeben, Items bei Nichtwissen auszulassen und somit das Risiko von Minuspunkten zu umgehen (Holmes, 2002, S. 15). Bei diesem Vorgehen besteht jedoch die Gefahr, dass risikovermeidende Personen durch eine solche Korrektur benachteiligt werden, wenn sie instruiert werden, nicht zu raten und bei Nichtwissen Items auszulassen. Es konnte gezeigt werden, dass risikovermeidende Personen dazu tendieren Items auszulassen, wenn sie sich bezüglich der Antwort nicht sicher genug sind. Die Benachteiligung entsteht dadurch, dass durch Raten auf Basis von Teilwissen ein höherer Score erwartet werden kann, als wenn bei Unsicherheit Items ausgelassen werden (Slakter, 1968). Bruno (1993) kritisiert, dass während der Score durch das Multiple-Choice Format auf Grund des Rateeffekts nach oben hin verzerrt wird, es durch die nachträgliche Ratekorrektur zu einer Verzerrung nach unten hin kommt, da zufälliges uninformiertes Raten von vornherein angenommen wird (zitiert nach Holmes 2002, S. 16). Schaefer (1976) diskutiert in diesen Zusammenhang auch die der Ratekorrektur zugrunde liegende Annahme, dass die Testperson bei Items, deren Lösung sie nicht definitiv richtig wissen, rein zufällig raten. Somit müsste jede Antwortmöglichkeit die gleiche Wahrscheinlichkeit haben, geraten zu werden. Voraussetzung hierfür ist, dass die Testperson über kein partielles Wissen verfügt. Kann die Testperson nämlich eine oder mehrere Antwortmöglichkeiten ausschließen, so verbessert sich ihr Score erheblich und wird durch die Ratekorrektur nach oben hin verzerrt. Lässt sich die Testperson hingegen von Fehlinformationen leiten, kann es, wie auch von Bruno (1993, zitiert nach Holmes, S.16) kritisiert, zu einer Verzerrung des Scores nach unten kommen.

Zu den auf der Item-Response-Theorie beruhenden Ansätzen dem Rateeffekt entgegenzuwirken, gehören die Rateparameter im 3-PL-Modell (vgl. Birnbaum, 1968), das „Difficulty-plus guessing“-Modell (vgl. Kubinger & Draxler, 2006, 2007) sowie die „Person-fit-Indizes“ (vgl. Ponocny & Klauer, 2002).

Der „Person-fit Index“ bietet im Rahmen von probabilistischen Testmodellen die Möglichkeit Personen zu identifizieren, deren erzielter Score keine adäquate Abbildung ihrer Fähigkeiten darstellt (Meijer, 2003). In der Diplomarbeitsstudie von Teubenbacher (2009) konnten erste Hinweise darauf gefunden werden, dass diese Statistik ein geeignetes Instrument zur Identifizierung von ratenden Personen darstellt. Das 3-PL-Modell und das „Difficulty-plus guessing“ Modell werden von Kubinger (2009) als Möglichkeiten beschrieben, den Rateeffekt testtheoretisch in den Griff zu bekommen. Hierfür wird der gesuchte Personenparameter unter Berücksichtigung der itemspezifischen Rateparameter geschätzt. Die Chance, erfolgreich zu raten, wird somit in den Testwert miteinkalkuliert und es kommt zu einer fairen Verrechnung, vorausgesetzt, für alle Testpersonen gelten dieselben Rateparameter. Dies ist jedoch schwer überprüfbar und es ist darüber hinaus anzunehmen, dass Personen ein unterschiedliches Rate- bzw. Risikoverhalten aufweisen. Kubinger (2009) verweist darauf, dass die beiden Modelle, wenn sie gelten, zwar Verrechnungsfairness garantieren, jedoch nicht individuelles Glück oder Pech beim Raten berücksichtigen können. Dieses schlägt sich jedoch in der faktischen Testleistung der Person nieder.

„*Damit scheint der bessere Zugang, psychologische, nämlich (neue) inhaltliche oder formal-gestalterische Mittel zu finden, um den aufgezeigten Problemen des Rateeffekts beim Multiple-Choice Format zu begegnen* (Kubinger 2009, S. 137).

1.2 Partielles Wissen

Bei all den bisher vorgestellten formal-gestalterischen Strategien, um dem Rateeffekt entgegen zu wirken, wird der Testperson nicht die Möglichkeit gegeben, partielles Wissen anzugeben. Vielmehr wird sie dazu ermuntert, auch bei Nichtwissen bzw. unsicherem Wissen, zu raten bzw. das Item auszulassen oder sich für eine neutrale Kategorie, wie „ich weiß die Antwort nicht“, zu entscheiden. Diagnostische Information kann durch dieses Vorgehen verloren gehen.

Echternacht formuliert in diesem Kontext folgendes:

[…] *knowledge is neither a dichotomous nor a trichotomous affair, which traditional multiple choice tests seem to imply, but it is continuous in the sense that there are varying degrees of knowledge.“* (Echternacht, 1972, S. 218).

Nach Mondak (2001) lassen sich voll informierte, teil informierte, falsch informierte und uninformierte Personen unterscheiden. Diese sind jedoch aufgrund von Tests im Multiple-Choice Format nicht immer differenzierbar. Völlig ahnungslose Personen lassen sich nicht von solchen unterscheiden, die zumindest Teilwissen haben und zum Beispiel zwei Antwortmöglichkeiten ganz sicher ausschließen können.

Nadeau und Niemi (1995) beschreiben in diesem Zusammenhang zwei Formen des Ratens, das reine Raten *(wild guessing*) und das Raten auf der Grundlage partiellen Wissens (*educated guessing*).

Nach Schaefer (1976) scheint es bei Vorgabe eines Leistungstests im Multiple-Choice Format sinnvoll, fünf verschiedene Fälle von Wissen der Testperson zu unterscheiden:

1. Vollständiges Wissen: Die Testperson kann die richtige Antwortmöglichkeit mit Sicherheit identifizieren.
2. Unkenntnis: Die Testperson besitzt bezüglich aller zu einem Item dazugehörigen Antwortmöglichkeiten keinerlei Information.
3. Partielle Information I: Die Testperson kann zwar keine Antwortmöglichkeit definitiv eliminieren, sie kann den einzelnen Antwortmöglichkeiten jedoch unterschiedliche Plausibilitätsgrade zuordnen.
4. Partielle Information II: Die Testperson kann eine bestimmte Anzahl an Distraktoren identifizieren und eliminieren.

5. Fehlinformationen: Die Testperson identifiziert mit subjektiver Sicherheit eine falsche Antwortmöglichkeit.

Im folgenden Kapitel wird ein Verfahren beschrieben, welches die Möglichkeit bietet, verschiedene Stufen von Wissen zu erfassen, indem der Testperson im Fall von unsicherem Wissen die Möglichkeit geboten wird, dieses anzugeben, anstatt zu raten.

2. Das Multiple-Evaluations Format

Das Multiple-Evaluations Format ist ein bisher in der Psychologischen Diagnostik kaum eingesetztes und wenig erforschtes Antwortformat. In den folgenden Unterkapiteln wird das Format im Detail vorgestellt und mögliche Vor- und Nachteile diskutiert.

2.1 Probabilistisches Messen

Wie auch beim Multiple-Choice Format wird beim probabilistischen Messen der Testperson pro Item eine gewisse Anzahl an Antwortmöglichkeiten vorgegeben. Sie wird jedoch nicht dazu aufgefordert, sich für eine Antwortmöglichkeit zu entscheiden, sondern hat 100 Prozent zu Verfügung, die sie auf die Antwortmöglichkeiten verteilen kann. Dabei sollen die auf eine Antwortmöglichkeit gesetzten Prozent angeben, für wie wahrscheinlich die Testperson diese Antwortmöglichkeit für richtig hält. Scheinen der Person zum Beispiel zwei der ihr vorgegebenen Antwortmöglichkeiten gleich plausibel und kann sie die anderen Antwortmöglichkeiten ausschließen, entspricht eine Vergabe von je 50 Prozent auf die plausiblen Antwortmöglichkeiten dem Wissenstand der Person.

Das Responseformat, also die Art wie die Testperson auf ein Item zu antworten hat, zeichnet sich beim probabilistischen Messen also nicht mehr durch das Ankreuzen bzw. Erkenntlichmachen der vermeintlich richtigen Antwortmöglichkeit aus, sondern sieht eine Vergabe von Wahrscheinlichkeiten in Form von Prozenten vor (vgl. Schafer, 1976).

Auch das Scoring erfolgt zwingenderweise anders als beim herkömmlichen Multiple-Choice Format. Die Verrechnung einer richtigen Antwort mit einem Punkt und einer falschen mit null Punkten ist nicht mehr in jedem Fall möglich.

Eine in der Literatur häufig diskutierte Verrechnungsmöglichkeit ist die Lineare Auswertung (LA). Bei dieser entspricht der Item-Score der Wahrscheinlichkeit, die die Testperson auf die richtige Antwort setzt (Rippey, 1970). Nach Dirkzwager (2003) ist bei einer solchen Verrechnungsart der zu erwartende Score maximal, wenn die Testperson immer 100 Prozent auf die ihr am plausibelsten erscheinende Antwortmöglichkeit setzt, anstatt ihr partielles Wissen in Form von Wahrscheinlichkeiten wiederzugeben.

Einige Autoren (Shuhford, Albert & Massengill, 1966; Holmes, 2002; Dirkzwager, 2003) sprechen sich deshalb für eine logarithmische Verrechnung aus, auf welcher auch das Multiple-Evaluations Format basiert. Dieses wird in den folgenden Kapiteln näher erläutert.

2.2 Das Multiple-Evaluations Format als Spezialfall des probabilistischen Messens

Das Multiple-Evaluations Format (ME Format) ist ein Spezialfall des probabilistischen Messens und basiert auf einer logarithmischen Verrechnung.

Wie auch bei der Linearen Auswertung wird bei der Berechnung des Item-Scores nur die Wahrscheinlichkeit, die die Testperson auf die tatsächlich richtige Antwortmöglichkeit setzt, berücksichtigt. Im Gegensatz zur Linearen Auswertung kann sich bei der logarithmische Verrechnung ein negativer Item-Score ergeben, wenn die Testperson einen zu geringen Prozentsatz auf die richtige Antwortmöglichkeit setzt. Dies führt dazu, dass der Item-Score einer Person dadurch maximiert werden kann, dass sie im Fall von Unsicherheit nicht versucht, die richtige Antwort zu erraten und damit das Risiko eingeht, Strafpunkte (negativer Item-Score) zu erhalten, sondern ihr Teilwissen korrekt und unverfälscht wiedergibt (vgl. Holmes, 2002). Die gesetzten Wahrscheinlichkeiten der Person müssen dafür der tatsächlichen subjektiven Antwortsicherheit der Person entsprechen (vgl. Shuhford, Massengil & Albert, 1966; Holmes, 2002, Dirkzwager, 2003).

Nach Dirkzwager setzt dies jedoch voraus, dass die Testpersonen in der Lage sind, ihr Wissen realistisch einzuschätzen. Es kann davon ausgegangen werden, dass es sowohl Personen gibt, die dazu tendieren ihr Wissen zu unterschätzen als auch solche, die dazu tendieren es zu überschätzen. Beide Phänomene können die Reliabilität sowie die Validität eines Tests im Multiplen-Evaluations Format gefährden. Dirkzwager spricht sich deshalb für eine ausreichende Übungszeit mit dem Antwortformat sowie für eine Vorgabe am Computer mit entsprechenden direkten Feedbackmöglichkeiten zu jedem Item aus (Dirkzwager, 1996, 2003; Holmes, 2002).

Im folgenden Kapitel wird die logarithmische Verrechnung, die das Multiple-Evaluations Format auszeichnet, im Detail beschrieben.

2.3 Logarithmische Verrechnung

Shuford et al. schlagen bereits im Jahr 1966 eine Verrechnungsregel vor, die auf einer logarithmischen Auswertung beruht. Basierend auf dem Modell von Shuford et al. schlägt Dirkzwager (2003) eine verbesserte logarithmische Verrechnungsregel vor.

Der Score für das Item i (s(i)) wird dabei unter Berücksichtigung eines Toleranzparameters (t) aus der Wahrscheinlichkeit, welche die Testperson auf die richtige Antwortmöglichkeit setzt (rc_i) und der Anzahl der Antwortmöglichkeiten pro Item (k) errechnet (Dirkzwager, 2003, S. 339).

$$s(i) = \frac{ln[(1-tk)\times r_c(i)+t]+ln(k)}{ln(1-tk+t)+ ln(k)} \quad (1)$$

Setzt die Testperson 100 Prozent auf die richtige Antwortmöglichkeit erhält sie einen Punkt. Setzt sie 0 Prozent auf die richtige Antwortmöglichkeit erhält sie die maximale Zahl an Minuspunkten pro Item (T).

Die Wahl des Toleranzparameters (t) bestimmt die maximale Zahl an möglichen Minuspunkten pro Item (T). Diese Zahl (T) entspricht der Anzahl an Items, die mit 100 Prozent auf der richtigen Antwortmöglichkeit beantwortet werden müssen, um eine völlig falsch beantwortete Frage (0 Prozent auf der richtigen Antwortmöglichkeit) zu kompensieren. Der Toleranzparameter limitiert also die Minuspunkte für Personen, die die richtige Antwort als sehr unwahrscheinlich einschätzen. Diese Fehleinschätzung kann darauf zurückzuführen sein, dass die Person nicht ihr wahres Teilwissen wiedergibt, sondern „wild rät“ oder, dass Fehlinformationen über die richtige Antwortmöglichkeit bestehen (Holmes, 2002 S. 43- 44).

Tabelle 1: Toleranzparameter (Holmes, 2002, S.44)

	k= 2	***k=3***	***k=4***	***k=5***
ME_1 (T= 1)	0, 4999	0, 1667	0, 0833	0, 0500
ME_2 (T=2)	0, 1910	0, 0447	0, 0174	0, 0086
ME_3 (T=3)	0, 0804	0, 0134	0, 0041	0, 0016
ME_4 (T=4)	0, 0362	0, 0043	0, 0010	0, 0003

Tabelle 1 zeigt die Toleranzparameter (t), die gewählt werden müssen um bei einer gegebenen Anzahl von Antwortmöglichkeiten (k) den entsprechenden Wert T (maximale Anzahl an Minuspunkten pro Item) zu erhalten. Je nach eingesetztem Toleranzparameter (entsprechend den T-Werten 1 bis 4) ergeben sich vier verschiedene Multiple-Evaluations Score-Berechnungsarten, die mit den Abkürzungen ME_1 bis ME_4 bezeichnet werden. In dieser Arbeit werden die logarithmischen Score-Verrechnungsarten ME_1 und ME_3 sowie die Lineare Auswertung (LA) zum Vergleich herangezogen. In der folgenden Tabelle (2) werden ME_1 und ME_3 sowie LA beispielhaft miteinander verglichen.

Tabelle 2:Vergleich der Score-Möglichkeiten ME_1, ME_3 und der linearen Auswertung

% auf die richtige Antwort	**LA**	**ME_1**	**ME_3**
0 %	0	- 1	- 3
10%	0,1	- 0,34	- 0,43
20 %	0,2	0	0
50%	0,5	0,54	0,57
80%	0,8	o,85	0,86
100%	1	1	1

2.4 Vor- und Nachteile des Multiplen-Evaluations Formats

Das Multiple-Evaluations Format stellt eine mögliche Strategie dar, dem Rateeffekt bei einem gebundenen Antwortformat entgegenzuwirken. Dies geschieht einerseits durch das spezielle Responseformat, bei dem der Testperson durch die Vergabe von Wahrscheinlichkeiten, entsprechend ihrer subjektiven Antwortsicherheit, die Möglichkeit gegeben wird, partielles Wissen ehrlich wiederzugeben. Andererseits wird durch die spezielle Art des Scorings, bei der „falsches" Raten in Form von Minuspunkten „bestraft" wird, ebenfalls eine Reduktion des Rateeffekts erwartet (vgl. Holmes, 2002).

Es wird davon ausgegangen, dass jene Personen von einer logarithmischen Auswertung profitieren, die ihr Wissen realistisch einschätzen und wiedergeben und dadurch ihre Punkte maximieren (vgl. Shuhford et al, 1966; Holmes, 2002; Dirkzwager, 2003). Personen, die tendenziell hohe Wahrscheinlichkeiten vergeben, schneiden immer dann durch eine logarithmische Auswertung auf Grund der Strafpunkte schlechter ab, wenn sie ihr Wissen nicht realistisch einschätzen und hohe Prozente auf Distraktoren setzen. Schätzen sie ihr Wissen hingegen realistisch ein und geben dieses auch so wieder, schneiden sie sowohl bei der logarithmischen als auch bei der linearen Auswertung gleichermaßen gut ab.

Bezüglich der Reliabilität sowie der Validität im Vergleich zu Tests im herkömmlichen Multiple-Choice Antwortformat liegen bereits seit den 70er Jahren erste Ergebnisse vor.

Schaefer (1976) gab einer Gruppe von Studenten über ein Semester verteilt in insgesamt sechs Sitzungen einen Test im Multiplen-Evaluations Format vor und wertete diese sowohl mit einer konventionellen Methode als auch logarithmisch aus. Ein Vergleich der Reliabilität der beiden Scoring Verfahren zeigte einen geringen Reliabilitätsgewinn durch die logarithmische Auswertung. Dies kann jedoch laut Autor sowohl auf den geringen Stichprobenumfang, nur 21 Testpersonen nahmen an allen sechs Sitzungen teil, als auch auf den zu geringen Schwierigkeitsgrad der Fragen, die Testpersonen konnten im Durchschnitt 79, 9% der Fragen mit 100 Prozent Wahrscheinlichkeit richtig beantworten, zurückgeführt werden.

Holmes fand 2002 eine signifikante Verbesserung der Reliabilität durch die Vorgabe eines Tests im Multiplen-Evaluations Format. Er konnte nachweisen, dass ein Test im Multiple-Choice Antwortformat fünf Mal so viele Items bräuchte, um dieselbe Reliabilität zu erreichen, wie ein Test im Multiplen-Evaluations Format (zitiert nach Dirkzwager, 2003).

Musch stellte 2009 ebenfalls Ergebnisse vor, die für eine höhere Reliabilität durch die Vorgabe des Multiplen-Evaluations Formats sprechen (Musch 2009, unveröffentlichte Ergebnisse).

Hambleton, Roberts und Traub (1970) führten eine Untersuchung mit 211 Student/innen durch, wobei einem Teil von ihnen eine Prüfung im Multiplen-Evaluations Format vorgegeben wurde und dem anderen Teil eine Prüfung im herkömmlichen Multiple-Choice Format. Eine Schätzung der Reliabilitäten mit der Split-Half Technik konnten zeigen, dass der Test im Multiplen-Evaluations Format eine geringere Reliabilität aufweist, als der Test im Multiple-Choice Format. Zwischen den Multiplen-Evaluations Scores und den Scores einer weiteren Testung am Semesterende, welche unter einer herkömmlichen Multiple-Choice Bedingung stattfand, fanden die Autoren höhere Korrelationen als zwischen den herkömmlichen Multiple-Choice Scores bei der ersten Testung und den Semesterendscores. Sie interpretieren diese Ergebnisse als Verbesserung der Validität durch die Vorgabe eines differenzierteren Antwortformates.

Bezüglich der Validität eines Tests im Multiplen-Evaluations Format wurden jedoch von einigen Autoren Zweifel geäußert. Sie vermuten, dass ein Einfluss von persönlichkeitsspezifischen Faktoren auf die Vergabe der Wahrscheinlichkeiten nicht ausgeschlossen werden kann (vgl. Schaefer, 1976).

Im folgenden Kapitel werden Untersuchungen zu möglichen Persönlichkeitseinflüssen auf das Antwortverhalten von Personen beim probabilistischen Messen vorgestellt und diskutiert.

3. Einflussfaktoren auf das Antwortverhalten beim probabilistischen Messen

Damit beim probabilistischen Messen valide Informationen über den Wissenstand der Testperson erhalten werden, müssen deren auf die einzelnen Antwortmöglichkeiten gesetzten Wahrscheinlichkeiten, ihrem tatsächlichen Wissen bzw. ihrer tatsächlichen Fähigkeit entsprechen. Sollten andere Merkmale wie die Persönlichkeit oder das Geschlecht der Testpersonen deren Antwortverhalten beeinflussen, kann der Einsatz eines solchen Tests als problematisch angesehen werden (Hansen, 1971).

3.1 Antwortverhalten beim probabilistischen Messen

Als ein Maß, um das Antwortverhalten einer Testperson auf Grundlage der von ihr vergebenen Prozentverteilung zu beschreiben, schlägt Hansen (1971) die durchschnittliche Itemsicherheit (C_T) vor. Sie kann bei einem Tests mit beliebig vielen Antwortmöglichkeiten (k) pro Item über alle Items (N) auf Grundlage der auf die einzelnen Antwortmöglichkeiten der Items gesetzten Wahrscheinlichkeiten (p_{ij}), nach folgender Formel berechnet werden:

$$C_T = \frac{1}{N}\sum_{j=1}^{N}\frac{k}{2\,(k-1)}\sum_{i-1}^{k}\left|\frac{1}{k} - p_{ij}\right| \qquad (2)$$

Bei einem Item mit k Antwortmöglichkeiten ist die durchschnittliche Itemsicherheit die Summe der absoluten Abweichungen von einer 1/k Gleichverteilung. Sie erreicht dementsprechend ihren minimalen Wert null, wenn die Testperson ihre Prozente gleichverteilt (1/k) und ihren maximalen Wert eins, wenn sie auf eine der Antwortmöglichkeiten 100 Prozent setzt, unabhängig davon, ob es sich um die richtige Antwortmöglichkeit handelt oder nicht (Hansen, 1971). Die durchschnittliche Itemsicherheit beschreibt damit das Ausmaß an Sicherheit einer Testperson in Bezug auf ihre Antworten.

Hansen (1971) konnte zeigen, dass Individuen beim probabilistischen Messen eine charakteristische Tendenz aufweisen, sicher oder unsicher zu antworten. Er konnte außerdem zeigen, dass diese Tendenz zwischen mehreren Testzeitpunkten relativ stabil bleibt. Diese Stabilität kann nicht vollständig auf die Stabilität des Wissens der Person zurückgeführt werden. Es konnte nur ein geringer Zusammenhang zwischen der individuellen Tendenz der Testpersonen in einer charakteristischen Form sicher oder unsicher zu antworten und ihrem Wissen gefunden werden.

3.2 Situative Einflüsse auf das Antwortverhalten

Sieber (1974) untersuchte an 40 Kollegstudent/innen situative Einflüsse auf das Antwortverhalten der Testpersonen bei Vorgabe eines probabilistischen Tests. Eine Gruppe von Student/innen wurde hierfür im Glauben gelassen, die Prüfung bei schlechtem Abschneiden nicht mehr wiederholen zu dürfen, während der anderen Gruppe mitgeteilt wurde, sie hätten eine Wiederholungsmöglichkeit. Er konnte zeigen, dass die Gruppe der Student/innen, welche die Testung als wichtiger ansah, da keine Wiederholungsmöglichkeit bestand, ihr Wissen tendenziell sicherer einschätzten.

Der Autor führt die vorliegenden Ergebnisse darauf zurück, dass unter dem hemmenden Einfluss eines hohen Angstlevels, welcher in besonders entscheidenden Testsituationen häufig auftritt, die Fähigkeit einer Person, subjektive Unsicherheiten korrekt wieder zu geben, abnehmen kann. Er schließt daraus, dass der Einsatz probabilistischer Tests in solchen Situationen möglicherweise nicht zielführend ist (Sieber, 1974).

3.3 Persönlichkeitseinflüsse auf das Antwortverhalten

Echternacht, Boldt und Sellman untersuchten 1972 den Zusammenhang von Persönlichkeitsvariablen und dem Abschneiden bei einem Test im Multiplen-Evaluations Format. Hierzu wurde einer Gruppe von Student/innen ein Test im Multiplen-Evaluations Format und einer anderen Gruppe ein Test in einem, dem herkömmlichen Multiple-Choice Format ähnlichen, Antwortformat vorgegeben. Beide Gruppen wurden gebeten, mehrere Persönlichkeitsfragebögen zu beantworten. Nach einer längeren Übungsphase wurden die Student/innen zu zwei Testzeitpunkten geprüft. Bei beiden Antwortformaten wurden für keine der erfassten Persönlichkeitsvariablen Zusammenhänge mit dem Score gefunden, welche über die zwei Testzeitpunkte repliziert werden konnten.

1971 zeigte Hansen einen positiven Zusammenhang zwischen der durchschnittlichen Itemsicherheit einer Person bei einem probabilistischen Test und ihrer Risikobereitschaft. Personen, die eine höhere Risikobereitschaft aufwiesen, tendierten dazu sich bezüglich ihrer Antworten sicherer zu sein, als es ihrem wahren Wissen entsprach.

Koehler (1974) untersuchte ebenfalls den Zusammenhang zwischen der durchschnittlichen Itemsicherheit einer Person und ihrer Risikobereitschaft. Zusätzlich zu den herkömmlichen Items eines verbalen Tests wurden den Testpersonen 7 Nonsensitems vorgegeben. Nonsensitems sind solche, bei denen die Frage grundsätzlich nicht beantwortbar ist. Basierend

auf der durchschnittlichen Itemsicherheit der Testperson bei diesen 7 Nonsensitems, postulierte Koehler (1974) ein Überkonfidenz Maß. Er konnte einen moderaten Zusammenhang zwischen diesem Überkonfidenz Maß und der Risikobereitschaft einer Person finden. Problematisch an dieser Untersuchung ist, dass sowohl die Risikobereitschaft als auch das Überkonfidenz Maß einer Person aus deren Antworten auf die vorgegebenen Nonsensitems berechnet wurden. Aufgrund dieses Vorgehens scheint es dem Autor fraglich, inwiefern der gefundene Zusammenhang zwischen Risikoverhalten und dem Überkonfidenz Maß echt ist.

Schaefer, Williams, Goodie und Campebell (2004) gaben in einer neueren Studie 100 Student/innen einen Wissenstest vor, bei dem sie pro Item zwei Antwortmöglichkeiten zur Auswahl hatten. Die Testpersonen wurden aufgefordert, sich für eine der beiden Antwortmöglichkeit zu entscheiden und ihre persönliche Antwortsicherheit auf einer Skala von 50 bis 100 Prozent anzugeben. Es wurden drei Maße berechnet. Die durchschnittliche Antwortsicherheit, die durchschnittliche Richtigkeit der Antworten und die Überkonfidenz, welche die Differenz zwischen der durchschnittlichen Antwortsicherheit und der durchschnittlichen Richtigkeit angibt. Zusätzlich wurden die Persönlichkeitsvariablen Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit erhoben. Zwischen Extraversion und der Überkonfidenz der Testpersonen konnte ein positiver Zusammenhang gefunden werden. Ebenso zeigten sich positive Zusammenhänge zwischen der durchschnittlichen Antwortsicherheit sowie der durchschnittlichen Richtigkeit und der Persönlichkeitsvariable Offenheit für Erfahrungen.

3.4 Einflüsse des Geschlechts auf das Antwortverhalten

Zu Geschlechtsunterschieden bezüglich der Einschätzungen der eigenen Leistungen liegen zahlreiche Studien vor, die darauf schließen lassen, dass Frauen dazu neigen ihre Leistung zu unterschätzen beziehungsweise, dass sie sich bei gleicher erbrachter Leistung schlechter einschätzen als Männer (Hannover & Bettge, 1993; Heatherington, Daubman, Bates, Ahn, Brown & Preston, 1993; Ehrlinger & Dunning, 2003).

Die meisten Studien basieren auf allgemeinen Konfidenzmaßen wie der Vorhersage von Noten oder der Einschätzung der eigenen Fähigkeit einen Test zu bestehen. Über Geschlechtsunterschiede bezüglich der Sicherheit ein Item richtig gelöst zu haben, liegen kaum Studien vor (Lundenberg, Fox & Puncochar, 1994).

Lundeberg, Fox und Puncochar (1994) untersuchten an 70 Männern und 181 Frauen, ob Differenzen zwischen deren Sicherheit, ein einzelnes Item bzw. eine einzelne Frage richtig beantwortet zu haben, vorliegen. Sie forderten eine Gruppe von Student/innen bei einer Multiple-Choice Prüfung auf, nach jedem Item anzugeben, wie sicher sie sich sind, dass die von ihnen gewählte Antwortmöglichkeit die richtige ist. Sowohl Frauen als auch Männer zeigen eine Tendenz ihre Leistungen zu überschätzen. Es zeigt sich jedoch, dass bei falsch beantworteten Fragen Frauen das Ergebnis realistischer einschätzen als Männer.

III. Empirischer Teil

4. Zielsetzung der Untersuchung

Die vorliegende Arbeit dient der Überprüfung möglicher Einflüsse auf das Antwortverhalten einer Person bei Vorgabe eines Tests im Multiplen-Evaluations Format. Es wird untersucht, ob es neben dem Wissen bzw. der Fähigkeit einer Person noch andere Einflüsse darauf gibt, wie sicher sich eine Person bezüglich ihres Wissens ist und wie realistisch sie ihr Wissen dabei einschätzt. Hierfür wird einerseits der Zusammenhang bestimmter Persönlichkeitsvariablen der Testpersonen mit ihrer Antwortsicherheit (durchschnittliche Itemsicherheit nach Hansen, 1971) untersucht. Andererseits wird überprüft, ob bestimmte Persönlichkeitsmerkmale einen Einfluss darauf haben, dass die Testpersonen von einer logarithmischen Auswertung (mit Strafpunkten) relativ zu einer linearen Auswertung (ohne Strafpunkte) profitieren. Zusätzlich werden in dieser Arbeit auch mögliche Geschlechtseinflüsse auf das Antwortverhalten untersucht.

4.1 Hypothesen

4.1.1 Hypothesen zu Persönlichkeitseinflüssen auf die Antwortsicherheit

H_1(a,b,c,d,): Es besteht ein signifikanter Zusammenhang zwischen der Persönlichkeitsvariablen Neurotizismus (a), Extraversion (b), Altruismus (c) sowie Gewissenhaftigkeit (d) und der durchschnittlichen Itemsicherheit einer Person.

Basierend auf der vorliegenden Literatur ergeben sich für die Persönlichkeitsvariablen Offenheit für Erfahrungen (Schaefer, Williams, Goodie & Campebell, 1994) und Risikobereitschaft (Hansen, 1971; Koehler, 1974) folgende gerichtete Hypothesen:

H_1(e): Es besteht ein positiver signifikanter Zusammenhang zwischen der Persönlichkeitsvariablen Offenheit für Erfahrungen und der durchschnittlichen Itemsicherheit einer Person.

H_1(f): Es besteht ein positiver signifikanter Zusammenhang zwischen der Persönlichkeitsvariablen Risikobereitschaft und der durchschnittlichen Itemsicherheit einer Person.

4.1.2 Hypothesen zu Geschlechtsunterschieden bezüglich der Antwortsicherheit

Auf Grundlage der von Lundeberg, Fox, und Puncochar (1994) gefundenen Ergebnisse ergibt sich zu den Geschlechtsunterschieden bezüglich der Antwortsicherheit folgende gerichtete Hypothese:

H_1(g): Männer sind signifikant antwortsicherer (höhere durchschnittliche Itemsicherheit) als Frauen.

4.1.3 Hypothesen zu Persönlichkeitseinflüssen auf das Ausmaß des Profites durch eine logarithmische Auswertung

Zur Prüfung möglicher Persönlichkeitseinflüsse auf die Tendenz einer Person, von einer logarithmischen Auswertung zu profitieren, ergeben sich folgende Hypothesen:

H_1(h, i, j, k, l, m): Es gibt einen signifikanten Zusammenhang zwischen der Persönlichkeitsvariablen Neurotizismus (h), Extraversion (i), Altruismus (j), Offenheit für Erfahrungen (k), Gewissenhaftigkeit (l) sowie Risikobereitschaft (m) und dem Ausmaß des Profits von einer logarithmischen relativ zu einer linearen Auswertung.

4.1.4 Hypothesen zu Geschlechtsunterschieden bezüglich des Ausmaßes des Profits von einer logarithmische Auswertung

H_1(n): Männer und Frauen unterscheiden sich signifikant bezüglich des Ausmaßes ihres Profits von einer logarithmischen relativ zu einer linearen Auswertung.

4.1.5 Hypothesen zum Einfluss der Antwortsicherheit einer Person auf das Ausmaß ihres Profits durch eine logarithmische Auswertung

In der vorliegenden Arbeit wird überprüft, inwiefern ein Zusammenhang zwischen der durchschnittlichen Itemsicherheit einer Person und ihrer Tendenz von einer logarithmischen Auswertung zu profitieren, besteht.

H_1(o): Es besteht ein signifikanter Zusammenhang zwischen der durchschnittlichen Itemsicherheit einer Person und dem Ausmaß ihres Profits von einer logarithmischen relativ zu einer linearen Auswertung.

5. Methode

Im folgenden Kapitel werden der Plan, die angewendeten Erhebungsinstrumente sowie die Durchführung der Untersuchung beschrieben.

5.1 Untersuchungsplan

Obwohl sich sowohl Dirkzwager (1996, 2003) als auch Holmes (2002) für eine Vorgabe von Tests im Multiplen-Evaluations Format am Computer aussprechen, wird aus ökonomischen Gründen eine Testung der Schüler/innen im Papier-Bleistift Format einer Computertestung vorgezogen.

Die Testung erfolgt an einem Wiener Gymnasium an Schüler/innen der 10ten, 11ten und 12ten Schulstufe, da angenommen werden kann, dass Schüler/innen dieser Altersstufe das Antwortformat auch ohne Computerrückmeldung entsprechend schnell reflektieren und sinnvoll anwenden können. Um eine Umsetzung im Papier-Bleistift Format gut durchführen zu können, werden eine klar formulierte, der Zielgruppe angepasste Instruktion sowie ausreichende Nachfrage- bzw. Nachbesprechungsmöglichkeiten angeboten. Anhand von drei Übungsfragen wird die, hinter dem Antwortformat stehende, Verrechnung erklärt.

Als Test zur Messung der verbalen Fähigkeiten wird der Untertest Satzergänzung aus dem I-S-T 2000R (Liepmann, Beauducel, Brocke & Amthauer, 2007) im Multiplen-Evaluations Format verwendet.

Zur Erfassung der Persönlichkeitsvariablen dient das Hamburger Persönlichkeitsinventar (Andresen, 2002).

Darüber hinaus werden noch folgende Daten zur Person erhoben: Alter, Geschlecht und Muttersprache.

Als Maß für die Antwortsicherheit der Testpersonen wird die durchschnittliche Itemsicherheit nach Hansen herangezogen.

Das Ausmaß des Profits von einer logarithmischen Auswertung (mit Strafpunkten) relativ zu einer linearen (ohne Strafpunkte) stellt eine weitere Variable dar und errechnet sich aus der Differenz zwischen dem z-transformierten Multiplen-Evaluations Score und dem z-transformierten Score der linearen Auswertung.

Alle nachfolgenden Analysen werden mit den Multiplen-Evaluations Scores ME_1 (maximal ein Minuspunkt pro Item) und ME_3 (maximal 3 Minuspunkte pro Item) durchgeführt (siehe Kapitel 2.3).

Zur Prüfung des Einflusses von Persönlichkeitsvariablen auf das Antwortverhalten dienen die Persönlichkeitsskalen des HPIs (Andresen, 2002) als unabhängige Variablen und die durchschnittliche Itemsicherheit (Hansen, 1971) bzw. das Ausmaß des Profits von einer logarithmischen Auswertung als abhängige Variable. Die Unterschiede zwischen Männern und Frauen hinsichtlich ihres Antwortverhaltens werden mit dem Geschlecht als unabhängige Variable und der durchschnittlichen Itemsicherheit bzw. dem Ausmaß des Profits von einer logarithmischen Auswertung als abhängige Variable überprüft.

5.2 Erhebungsinstrumente

5.2.1 Satzergänzung (I-S-T 2000R)

Der Intelligenz-Struktur-Test 2000R (Liepmann, Beauducel, Brocke & Amthauer, 2007) ist ein Intelligenztest, der Jugendlichen ab 15 Jahren vorgegeben werden kann und verbale, numerische und figurale Intelligenz sowie logisches Denken erfasst. Zusätzlich können verbale und figurale Merkaufgaben sowie ein Wissenstest vorgegeben werden.

Der für den empirischen Teil dieser Arbeit vorgegebene Untertest Satzergänzung erfasst neben zwei anderen Untertest die verbale Intelligenz. Jede Aufgabe besteht aus einem Satz, in dem ein Wort fehlt. Aus fünf vorgegebenen Wörtern soll jenes ausgewählt werden, das den Satz richtig vervollständigt. Der Untertest wird im Original in einem 5-kategorialen Multiple-Choice Format vorgegeben und setzt sich aus zwanzig Items zusammen (Liepmann, Beauducel, Brocke & Amthauer, 2007).

Dem eigentlichen Test gehen zwei Beispiel-Items voran, die wie folgt lauten:

Ein Kaninchen hat am meisten Ähnlichkeit mit einem (einer)…?

a)Katze b) Eichhörnchen c) Hasen d) Fuchs e) Igel

Das Gegenteil von Hoffnung ist…?

a)Trauer b) Verzweiflung c) Elend d) Liebe e) Hass

5.2.2 HPI

Das Hamburger Persönlichkeits-Inventar (Andresen, 2002) kann ab 16 Jahren vorgegeben werden und erfasst, neben den fünf Dimensionen nach dem Faktorenmodell (Costa und McCrae) den Faktor „Risiko-und Kampfbereitschaft, Suche nach Wettbewerb“. Da diesem Persönlichkeitsfaktor auf Grund der vorliegenden Literatur besonderes Interesse beigemessen werden kann, wird in dieser Arbeit das Hamburger Persönlichkeitsinventar anderen Persönlichkeitstest vorgezogen.

Im Folgenden werden alle sechs Skalen des HPI’s (Andresen, 2002) kurz beschrieben.

Skala N : Nervosität, Sensibilität und emotionale Instabilität

Diese Skala erfasst die klassische „Neurotizismus Dimension“, jedoch nicht im traditionell klinischen Sinn beziehungsweise primär psychopathologisch definiert. Die Skala N kann als Maß für psychosoziale Stressreagibiliät und Bereitschaft zu valenznegativen Affekten gesehen werden, vor allem im sozial-interaktiven und -evaluativen Kontext.

Skala E: Extraversion, Lebhaftigkeit und Kontaktfreude
Die Skala E beschreibt positiv-emotionale, sozial-interaktive Erlebnisbereitschaft, welche mit Unternehmungslust, Abwechslungsbedürfnis und Unterhaltungswünschen sowie aktiver Lebenslust einhergeht.

Skala O: Offenheit für Erfahrungen
Der Fokus der Skala O liegt auf der Fantasietätigkeit, Erlebnisoffenheit, kreativer Neigung und dem Interesse an musischen, weltanschaulichen sowie psychologischen Inhalten einer Person. Hohe Werte auf dieser Skala sprechen für einen kulturell offenen, unkonventionellen und zugleich nachdenklich-subjektorientierten Menschen.

Skala C: Kontrolliertheit und Normorientierung
Die Skala C erfasst Bereiche der Konsequenz, Rigidität und Selbstkontrolle sowie konservative Orientierungen. Hohe Werte in dieser Skala sprechen für ausgeprägte Ordentlichkeit, Pünktlichkeit, Genauigkeit sowie Moral, Strenge und Pflichtbewusstsein.

Skala A: Altruismus, Fürsorglichkeit und Hilfsbereitschaft
Die Skala A steht für prosoziale Orientierung, Zärtlichkeitsbedürfnis, Suche nach Harmonie und Friedfertigkeit sowie Selbstaufopferungstendenzen, Hingabefähigkeit und Empathie.

Skala R: Risiko- und Kampfbereitschaft, Suche nach Wettbewerb

Unter Risikobereitschaft verstehen die Autoren die riskierende und zugleich an „starken“ moralischen Werten orientierte Auseinandersetzung mit herausfordernden, gefährlichen oder schwierigen Umweltsituationen im weitesten Sinne (Andresen, 1995, S. 210). Die Skala beschreibt das kämpferische, sportive und wettbewerbsorientierte Verhaltensspektrum sowie Abenteuerlust, technische Interessen mit Risikoakzent und physisches Aktivitätsbedürfnis.

5.3 Durchführung der Untersuchung

Die Untersuchung fand in dem Zeitraum vom 1.12.2010 bis zum 6.12.2010 an einem Wiener Gymnasium im Rahmen des Psychologie- und Philosophieunterrichts statt. Es wurden in einer 6ten (10te Schulstufe), vier 7ten (11te Schulstufe) und vier 8ten Klassen (12te Schulstufe) Gruppentestungen durchgeführt. Die Testung dauerte jeweils eine Schulstunde (à 50 Minuten) und wurde durch Unterstützung der Direktion und der unterrichtenden Lehrerinnen ermöglicht. Den Lehrerinnen war es ein Anliegen, den Schüler/innen einen Einblick in das psychologisch wissenschaftliche Arbeiten zu geben.

5.3.1 Instruktion und Übungsbeispiele

Die Instruktion wurde persönlich und interaktiv anhand von Power-Point Folien (siehe Anhang 11.1) gestaltet. Zusätzlich wurde nach abgeschlossener Instruktion ein Instruktionsblatt mit den wichtigsten Informationen über das Multiple-Evaluations Format und dessen Auswertung bzw. Verrechnung an die Schüler/innen ausgeteilt (siehe Anhang 11.2).

Zum Einstieg wurden den Schüler/innen das Multiple-Choice Antwortformat und dessen Verrechnung näher gebracht, um die Unterschiede zum Multiplen-Evaluations Format zu verdeutlichen. Anhand von drei Übungsfragen mit je 5 Antwortmöglichkeiten (die Haupttestung erfolgte auch in einem 5-kategorialen Antwortformat) wurde gemeinsam geklärt, welches Vorgehen bei dem neuen Antwortformat sinnvoll wäre. Bei der ersten, sehr leicht zu lösenden Frage „Wie heißt die Hauptstadt von Österreich“ sollte gemeinsam erarbeitet werden, dass das Setzen von 100 Prozent bei absoluter Sicherheit zielführend ist. Das zweite Item „Welches ist das größte Iranische Gebirge“ sollte für den Großteil der Schüler/innen nicht lösbar sein. Dadurch sollte verdeutlicht werden, dass eine Gleichverteilung der 100 Prozent am sinnvollsten ist. Bei dem letzten Item „Welche ist die Hauptstadt von Australien“ wurde eine Frage gewählt, die die meisten Schüler/innen

zwischen mindestens zwei Antwortmöglichkeiten schwanken lassen sollte, um so ergänzend zu zeigen, welche verschiedene andere Möglichkeiten der Prozentverteilung es neben völliger Sicherheit (100 Prozent) und absolutem Nichtwissen (Gleichverteilung der Prozente) gibt. Im Anschluss an jede Frage wurde auch die entsprechende Punktevergabe besprochen. Die Verrechnung wurde somit auf einer Ebene transparent gemacht, die das System hinter dem Multiplen-Evaluations Format und der logarithmischen Verrechnung verständlich machen sollte, ohne jedoch mathematisch in die Tiefe zu gehen, um die Schüler/innen nicht unnötig zu verwirren.

Die Schüler/innen hatten während der gesamten Instruktion und auch während der Testung die Möglichkeit Fragen zu stellen und Missverständnisse zu klären.

Die meisten Schüler/innen nahmen das neue Antwortformat sehr gut auf und zogen intuitiv die richtigen Schlüsse in Bezug auf ein zielführendes Antwortverhalten. Unklarheiten schien es wenige zu geben und wenn, konnten diese während der Instruktionsphase behoben werden.

5.3.2 Testung

Danach folgte der Hauptteil der Testung, bei dem der Untertest Satzergänzung aus dem I-S-T 2000R (Liepmann, Beauducel, Brocke & Amthauer, 2007) vorgegeben wurde und die Schüler/innen gebeten wurden, entsprechend dem Multiple-Evaluations Format zu antworten. Um ein gemeinsames Bearbeiten der Fragen durch die Schüler/innen zu erschweren, wurden sowohl die Form A als auch die „Pseudoparallelform“ B (lediglich die Reihenfolge der Items ist vertauscht) des Untertests Satzergänzung vorgegeben.

Außerdem wurde den Schüler/innen das Hamburger Persönlichkeitsinventar (Andresen, 2002) vorgegeben.

Abschließend wurden die Schüler/innen ersucht, Daten zu ihrer Person (Alter, Geschlecht und Muttersprache) anzugeben.

5.4 Stichprobe

Das Alter der Schüler/innen liegt zwischen 15 und 19 Jahren und beträgt im Durchschnitt 16,81 Jahre.

Die Stichprobe setzt sich aus 54 Schülern (33,3%) und 108 Schülerinnen (66,7%) zusammen (siehe Abb. 5).

Rund 83% (135 Personen) der Stichprobe haben als Muttersprache Deutsch, knapp 17% (27 Personen) haben eine andere Muttersprache

6. Ergebnisse

Im folgenden Kapitel werden die Ergebnisse der statistischen Analyse dargestellt. Zu Beginn wird die Datenerfassung und Berechnung der Variablen beschrieben. Abschließend werden die Ergebnisse der Hypothesenprüfung dargeboten.

6.1 Datenerfassung und Berechnung der Variablen

Nach der Ausscheidung von sechs unvollständigen Fragebögen standen 162 Fragebögen zur Verfügung, die für die weiteren Berechnungen herangezogen wurden. Die Rohdaten aus dem Untertest Satzergänzung und dem Hamburger Persönlichkeitsinventar wurden in Excel 2007 erfasst. Anschließend wurden daraus der Score der Linearen Auswertung (LA) und die Scores der Multiplen-Evaluation (ME_1 und ME_3) sowie die durchschnittliche Itemsicherheit (C_T) berechnet. Die Rohscores der Persönlichkeitsskalen des HPIs wurden ebenfalls in Excel berechnet, die Umrechnung in die Standard Werte (SW) erfolgte direkt in PASW Statistics 18. Die Berechnung des Ausmaßes des Profits von ME_1 bzw. von ME_3 relativ zur LA erfolgte in PASW Statistics 18. Hierfür wurden die Variablen ME1, ME3 sowie LA vor der Differenzbildung (ME_1-LA, bzw. ME_3-LA) z-transformiert (M=0, SD= 1). Die statistische Analyse der Daten erfolgte in PASW Statistics 18.

6.2 Hypothesenprüfung

Im folgenden Kapitel werden die Ergebnisse der Hypothesenprüfungen dargestellt. Für die gesamte Analyse wurde ein Signifikanzniveau von $\alpha = 0{,}05$ festgelegt und eine Effektstärke von 1/3 (=0,33) als bedeutsam erachtet. Signifikante Ergebnisse werden zur besseren Übersicht fett unterlegt.

6.2.1 Ergebnisse zum Einfluss der Persönlichkeitsvariablen auf die durchschnittliche Itemsicherheit nach Hansen (C_T)

Um den Zusammenhang zwischen der durchschnittlichen Itemsicherheit (C_T) und den Persönlichkeitsskalen des HPIs zu berechnen, wurden Spearman Rangkorrelationen durchgeführt (siehe Tabelle 3). Es ist ersichtlich, dass zwischen dem C_T Wert und der Skala „Kontrolliertheit und Normorientierung“ (Gewissenhaftigkeit) ein signifikanter negativer Zusammenhang besteht ($r= -0{,}287$, $p<0{,}05$). Zwischen der Persönlichkeitsskala „Risiko-und

Kampfbereitschaft, Suche nach Wettbewerb“ und dem C_T Wert liegt ein signifikanter positiver Zusammenhang vor (r= 0,195, p<0,05).

Zwischen den Variablen „Nervosität, Sensibilität und emotionale Instabilität“ (Neurotizismus) „Extraversion, Lebhaftigkeit und Kontaktfreude“, „Offenheit für Erfahrungen“ sowie „Altruismus, Fürsorglichkeit und Hilfsbereitschaft“ und dem C_T Wert konnten keine signifikanten Zusammenhänge gefunden werden.

Tabelle 3: Rangkorrelationen des C_T-Werts mit den Persönlichkeitsvariablen des HPIs

		Skala N	***Skala E***	***Skala O***	***Skala C***	***Skala A***	***Skala R***
C_T	**r**	-0,050	-0,032	-0,016	**-0,287**	-0,107	**0,195**
	p	0,528	0,683	0,418	0,000	0,176	0,007

Legende : Skala N: „Nervosität, Sensibilität und emotionale Instabilität“ (Neurotizismus)
Skala E: „Extraversion, Lebhaftigkeit und Kontaktfreude“
Skala O: „Offenheit für Erfahrungen“
Skala C: „Kontrolliertheit und Normorientierung“ (Gewissenhaftigkeit)
Skala A: „Altruismus, Fürsorglichkeit und Hilfsbereitschaft“
Skala R: „Risiko-und Kampfbereitschaft, Suche nach Wettbewerb“
C_T Wert: „Durchschnittliche Itemsicherheit nach Hansen“

6.2.2 Ergebnisse zum Geschlechtsunterschied bezüglich der durchschnittlichen Itemsicherheit

Um Unterschiede zwischen Schülerinnen und Schülern hinsichtlich ihrer durchschnittlichen Itemsicherheit zu überprüfen, wurden Mittelwerte verglichen. In Anlehnung an Rasch, Kubinger und Moder (2011) wurden die Mittelwertunterschiede anhand des Welch-Tests für unabhängige Stichproben überprüft.

Die Prüfung mittels Welch-Tests ergab einen signifikanten Unterschied zwischen Schülern (M=10,06, SD=1,63) und Schülerinnen (M=9,17, SD=1,79) hinsichtlich ihres C_T Werts (t= 3,19, $p \leq 0,05$, d=0,53). Diese Ergebnisse zeigen, dass Schüler signifikant antwortsicherer sind als Schülerinnen. Es konnte ein als bedeutsam zu erachtender Unterschied gefunden werden (siehe Tabelle 4).

Tabelle 4: Mittelwertvergleiche von C_T zwischen Männern und Frauen (Welch-Test)

	Geschlecht	*N*	*Mittelwert*	*Standard-abweichung*	*t*	*p*	*Effektstärke*
C_T	männlich	54	10,06	1,63	3,19	**0,002**	0,53
	weiblich	108	9,17	1,79			

6.2.3 Ergebnisse zum Einfluss von Persönlichkeitsvariablen auf das Ausmaß des Profits von einer logarithmischen Auswertung

Die Zusammenhänge zwischen dem Ausmaß des Profits von einer logarithmischen relativ zu einer linearen Auswertung mit den Persönlichkeitsvariablen wurden mit Spearman Rangkorrelationen überprüft.

Wie man aus Tabelle 5 entnehmen kann, liegt zwischen dem Ausmaß des Profits von einer logarithmischen Auswertung und der Persönlichkeitsskala „Kontrolliertheit und Normorientierung“ (Gewissenhaftigkeit) sowohl bei einer ME_1 Auswertung (r=0,186, p ≤0,05) als auch bei einer ME_3 Auswertung (r=0,237, p≤ 0,05) ein signifikanter positiver Zusammenhang vor.

Zwischen der Skala „Risiko- und Kampfbereitschaft, Suche nach Wettbewerb“ und dem Ausmaß des Profits zeigt sich ein signifikanter negativer Zusammenhang sowohl für ME_1 (r= -0,247, p≤0,05) als auch für ME_3 (r=-0,238, p≤ 0,05).

Tabelle 5: Rangkorrelationen des Ausmaßes des Profits von ME_1 und ME_3 mit den Persönlichkeitsvariablen des HPIs

		Skala N	*Skala E*	*Skala O*	*Skala C*	*Skala A*	*Skala R*
„Profit“ ME1	**r**	0,022	-0,018	0,058	**0,186**	0,082	**-0,247**
	p	0,776	0,825	0,462	0,018	0,298	0,002
„Profit“ ME_3	**r**	0,034	-0,015	0,042	**0,237**	0,091	**-0,238**
	p	0,670	0,849	0,599	0,002	0,252	0,002

6.2.4 Ergebnisse zum Geschlechtsunterschied bezüglich des Ausmaßes des Profits von einer logarithmischen Auswertung

Sowohl bei der ME_1 Auswertung als auch bei der ME_3 Auswertung zeigen sich signifikante Unterschiede zwischen Schülern und Schülerinnen hinsichtlich ihrer Tendenz von einer logarithmischen Auswertung zu profitieren.

Bei der ME_1 Auswertung konnte, wie in Tabelle 6 ersichtlich, gezeigt werden, dass die Schülerinnen (M= 0,06, SD= 0,43) im Durchschnitt signifikant mehr von einer logarithmischen Auswertung profitieren, als die Schüler (M= -0,12, SD=0,46), (t= -2,43, $p \leq 0,05$, d=0,41).

Von einer ME_3 Auswertung profitieren Schülerinnen ebenfalls (M=-0,12, SD=0,75) signifikant mehr als Schüler (M=-0,25, SD= 0,76), (t= -2,94, $p \leq 0,05$, d=0,49).

Bei beiden Auswertungsmethoden konnte ein als bedeutsam zu erachtender Unterschied gefunden werden.

Tabelle 6: Mittelwertvergleiche des „Profits" von ME_1 und ME_3 zwischen Männern und Frauen (Welch-Test)

	Geschlecht	*N*	*Mittelwert*	*Standard-abweichung*	*t*	*p*	*Effektstärke*
„Profit" ME1	männlich	54	-0,12	0,46	**-2,43**	**0,017**	**0,41**
	weiblich	108	0,06	0,43			
„Profit" ME3	männlich	54	-0,25	0,76	**-2,94**	**0,004**	**0,49**
	weiblich	108	0,12	0,75			

6.2.5 Ergebnisse zum Einfluss der durchschnittlichen Itemsicherheit (C_T) auf das Ausmaß des Profits von einer logarithmischen Auswertung

Die Zusammenhänge zwischen der durchschnittlichen Itemsicherheit und dem Ausmaß des Profits von einer logarithmischen Auswertung (sowohl ME_1 als auch ME_3) sind in Tabelle 7 dargestellt.

Zwischen dem Ausmaß des Profits von einer logarithmischen Auswertung mit einem Minuspunkt und der durchschnittlichen Itemsicherheit zeigt sich ein signifikanter negativer Zusammenhang (r= 0,778, $p<0,05$).

Ein signifikanter negativer Zusammenhang mit der durchschnittlichen Itemsicherheit zeigt sich ebenfalls bei einer logarithmischen Auswertung mit maximal 3 Minuspunkten pro Item ($r= 0{,}872$, $p<0{,}05$).

Tabelle 7.: Rangkorrelationen der durchschnittlichen Itemsicherheit mit dem Ausmaß des Profits von einer logarithmischen Auswertung

		CT
„Profit" ME_1	**r**	**0,778**
	p	0,000
„Profit" ME_3	**r**	**0,872**
	p	0,000

7. Diskussion der Ergebnisse und Ausblick

Im folgenden Kapitel werden die Ergebnisse der Testung interpretiert und diskutiert sowie eine kritische Auseinandersetzung mit der Untersuchung und ein Ausblick für weitere mögliche Studien zum Antwortformat der Multiplen Evaluation geboten.

7.1 Diskussion der Ergebnisse

Die vorliegende Arbeit dient der Überprüfung von Persönlichkeits- und Geschlechtseinflüssen auf das Antwortverhalten im Multiplen-Evaluations Format.

Hierfür wurde überprüft, inwiefern Persönlichkeitsvariablen und das Geschlecht einen Einfluss auf die Antwortsicherheit einer Person bei einer Testung im Multiplen-Evaluations Format haben. Als ein Maß für die Antwortsicherheit wurde die durchschnittliche Itemsicherheit nach Hansen verwendet (siehe Kapitel 3.2).

Ferner wurde überprüft, inwiefern Persönlichkeitsmerkmale bzw. das Geschlecht sowie die Antwortsicherheit einen Einfluss darauf haben, ob Personen von einer logarithmischen Auswertung profitieren. Die Tendenz einer Person bei einer Testung von einer logarithmischen Auswertung mit Strafpunkten relativ zu einer linearen zu profitieren, kann dahin gehend interpretiert werden, dass Personen ein Antwortformat der Multiplen-Evaluation dann optimal nutzen, wenn sie ihr Wissen realistisch einschätzen und wiedergeben und dadurch ihre Punkte maximieren.

7.1.1 Diskussion der Ergebnisse zu Persönlichkeitseinflüssen auf das Antwortverhalten

In dieser Arbeit zeigt sich, wie auf Grundlage der vorliegenden Literatur (Hansen 1971; Koehler 1974) zu erwarten, ein signifikanter, wenn gleich geringer positiver Zusammenhang zwischen der Risikobereitschaft und der durchschnittlichen Itemsicherheit einer Person. Risikobereitere Schüler und Schülerinnen waren in ihrem Antwortverhalten tendenziell sicherer. Zwischen der Persönlichkeitsvariable Risikobereitschaft und dem Ausmaß des Profits von einer logarithmischen Auswertung, konnte ein ebenfalls signifikanter, wenn auch geringer negativer Zusammenhang gefunden werden. Personen mit einer geringer ausgeprägten Risikobereitschaft profitieren überdurchschnittlich von dem neuen Antwortformat.

Wie die statistischen Auswertungen in Kapitel 6 belegen, liegt zwischen der Persönlichkeitsvariable Gewissenhaftigkeit und der durchschnittlichen Itemsicherheit ein

signifikanter, wenn gleich geringer negativer Zusammenhang vor. Zwischen der Persönlichkeitsvariablen Gewissenhaftigkeit und dem Ausmaß des Profits von einer logarithmischen Auswertung zeigte sich ein ebenfalls signifikanter, wenn auch geringer positiver Zusammenhang.

Der von Schaefer, Williams, Goodie und Campebell (2004) gefundene, positive Zusammenhang zwischen der Antwortsicherheit und der Persönlichkeitsvariable Offenheit für Erfahrungen konnte in der vorliegenden Arbeit nicht bestätigt werden.

Auf Grund der vorliegenden geringen statistischen Zusammenhänge mit den Persönlichkeitsvariablen Risikobereitschaft und Gewissenhaftigkeit sowie den nicht vorhanden Zusammenhängen mit den anderen Persönlichkeitsvariablen des HPIs kann diese Arbeit nicht bestätigen, dass es für die Praxis von Testungen im Multiple-Evaluations Format relevante Einflüsse der Persönlichkeit auf das Antwortverhalten gibt.

7.1.2 Diskussion der Ergebnisse zum Einfluss des Geschlechts auf das Antwortverhalten

Basierend auf den von Lundeberg, Fox, und Puncochar (1994) gefundenen Ergebnissen wurde auch in dieser Arbeit überprüft, ob Männer signifikant antwortsicherer sind als Frauen. Die Hypothese konnte bestätigt und der gefundene Effekt als praktisch bedeutsam erachtet werden. Die Ergebnisse lassen sich dahingehend interpretieren, dass Frauen eine generelle Tendenz aufweisen, sich in ihrer Leistung weniger sicher einzuschätzen als Männer (Hannover und Bettge, 1993; Heatherington, Daubman, Bates, Ahn, Brown & Preston, 1993; Ehrlinger & Dunning, 2003).

Bezogen auf das Ausmaß des Profits von einer logarithmischen Auswertung relativ zu einer linearen konnten ebenfalls signifikante Geschlechtsunterschiede gefunden werden. Frauen profitieren signifikant mehr von einer logarithmischen Auswertung als Männer. Dies deutet darauf hin, dass Frauen ihr Wissen realistischer einschätzen und eher bereit sind ihre Unsicherheiten offen dar zu legen.

Auf Grund der vorliegenden Ergebnisse kann also davon ausgegangen werden, dass Frauen durch eine logarithmische Verrechnung mit Strafpunkten bevorzugt werden. Männer werden hingegen tendenziell von einer logarithmischen Auswertung benachteiligt, da sie entweder ihr Wissen nicht realistisch einschätzen oder nicht realistisch wiedergeben.

7.1.3 Diskussion der Ergebnisse zum Einfluss der Antwortsicherheit auf das Ausmaß des Profits von einer logarithmischen Auswertung

Zwischen der durchschnittlichen Itemsicherheit und dem Ausmaß des Profits von einer logarithmischen Auswertung konnte ein hoher signifikant negativer Zusammenhang gefunden werden. Antwortsicherere Personen profitieren tendenziell weniger von einer logarithmischen Auswertung.

Verfügen Personen über ein moderates Wissen, so werden sie, wenn sie zu antwortsicher sind, durch die logarithmische Auswertung bestraft. Schätzen sie ihr Teilwissen realistisch ein und geben dies auch so wieder, profitieren sie von dieser Art der Auswertung. Verfügen antwortsichere Personen über ein sehr hohes Wissen bzw. sind die ihnen vorgegebenen Items zu einfach, werden sie durch die verschiedenen Verrechnungsarten kaum belohnt oder bestraft, da antwortsichere Personen nur dann Strafpunkte erhalten, wenn sie die hohen Prozentsätze auf falsche Antwortmöglichkeiten (Distraktoren) setzen.

Der gefundene hohe negative Zusammenhang zwischen der Antwortsicherheit und dem Ausmaß des Profits von einer logarithmischen Auswertung bestätigt daher die von Hansen (1971) beobachteten Ergebnisse und lässt darauf schließen, dass sich die Antwortsicherheit der Testpersonen nur zu einem geringen Teil auf deren Wissen gründet.

7.2 Ausblick

Im Ausblick dieser Arbeit soll diskutiert werden, wie weitere Untersuchungen gestaltet werden können, um Persönlichkeits- und Geschlechtseinflüsse auf das Antwortverhalten besser zu überprüfen. Ergänzend sollen abschließend Möglichkeiten beschrieben werden, eine Testung im Multiplen-Evaluationsformat dahin gehend zu optimieren, dass Persönlichkeits- und Geschlechtseinflüsse bestmöglich reduziert werden.

Aus ökonomischen Gründen weist der Untersuchungsplan der vorliegenden Diplomarbeit einige Defizite auf. Vor allem die einmalige Testung mit dem neuen Antwortformat muss als problematisch betrachtet werden. Es wäre wünschenswert, den Testpersonen einen mehrmaligen Kontakt mit dem für sie völlig neuen Antwortformat zu ermöglichen, um sicher zu gehen, dass die Testpersonen das Format und die dazugehörige Auswertung ausreichend verstehen. Des Weiteren könnte durch eine mehrmalige Testung auch überprüft werden, wie stabil die Einflüsse von Persönlichkeit und Geschlecht auf das Antwortverhalten sind.

Außerdem wäre es sinnvoll weitere Untersuchungen an einer anderen Stichprobenpopulation als der der Schüler/innen durchzuführen. Schüler/innen sind möglicherweise zu sehr in einem System verankert, in dem, unabhängig vom eingesetzten Antwortformat, das Geben einer „richtigen“ Antwort als erwartet angenommen wird.

Ferner wäre es interessant zu überprüfen, ob es in einer realen Testsituation, bei denen die Ergebnisse tatsächliche Konsequenzen für die Schüler/innen nach sich ziehen, zu anderen Ergebnissen kommt. Nach Sieber (1974) kann davon ausgegangen werden, dass Personen in einer solchen Situation eher dazu neigen, ihr Wissen zu überschätzen.

In weiteren Studien könnte es lohnend sein, anstelle eines Persönlichkeitsfragebogens bestimmte Persönlichkeitsvariablen, wie zum Beispiel Risikobereitschaft, mit einem Objektiven Persönlichkeitstest zu erfassen.

Der Versuch dieser Arbeit, das Antwortverhalten einer Person bei einer Multiplen-Evaluations Testung durch deren durchschnittliche Itemsicherheit zu beschreiben, kann dahingehend kritisiert werden, dass diese sowohl von Persönlichkeitsfaktoren als auch vom Wissen beeinflusst wird. Das Ausmaß dieser beiden Einflüsse ist jedoch nicht bekannt und variiert von Person zu Person. Wenn zum Beispiel für eine Person der Großteil der Items einfach zu beantworten ist, da sie über ausreichend Wissen verfügt, lässt sich das Vorliegen einer hohen durchschnittlichen Itemsicherheit vorrangig durch das vorhandene Wissen

erklären. Ein möglicher Ansatz wäre, wie bereits bei Koehler (1974), mit Nonsensitems, welche grundsätzlich nicht beantwortbar sind, zu arbeiten. Bei Vorgabe solcher Items hat das Wissen der Testperson keinen Einfluss auf deren Antwortsicherheit. Es gilt zu überprüfen, ob durch ein solches Vorgehen andere bzw. stärkere Zusammenhänge zwischen der Antwortsicherheit einer Person und Persönlichkeitsvariablen gefunden werden können.

Um Persönlichkeits- und Geschlechtseinflüsse bei Vorgabe eines Tests im Multiplen-Evaluations Format zu reduzieren, bietet sich die Vorgabe von Übungsitems an, die der eigentlichen Testung vorausgehen. Anhand dieser Items ließe sich überprüfen, inwiefern das Multiple-Evaluations Format von den Testpersonen verstanden wurde. Dabei würde sich eine Testung am Computer besonders anbieten, da diese ein direktes, der Testperson angepasstes, Feedback ermöglicht. Es kann davon ausgegangen werden, dass Testpersonen erst lernen müssen, ihr Wissen in einer solchen Testsituation realistisch einzuschätzen und dass dies ein ausreichendes Training mit dem Antwortformat voraussetzt (vgl. Dirkzwager, 1996; Holmes 2002). Ob sich durch eine in dieser Art optimierte Testsituation (ausreichend Feedback und Übungsmöglichkeiten) Persönlichkeits- und Geschlechtseinflüsse aufheben lassen, kann in Folgestudien überprüft werden.

8. Zusammenfassung

Das Multiple-Choice Antwortformat dominiert vor allem auf Grund seiner ökonomischen Durchführung und Auswertung das Feld der modernen Testkonstruktion. Die fehlende Möglichkeit durch das Antwortformat Teilwissen zu erfassen, da die Testpersonen sich auch dann, wenn sie über ein solches verfügen, für eine Antwortmöglichkeit entscheiden und somit raten müssen, kann als ein Nachteil dieses Antwortformats gesehen werden.

Probabilistisches Messen bietet die Möglichkeit, bei Vorgabe eines gebundenen Antwortformats partielles Wissen zu erfassen, in dem die Testperson pro Item 100 Prozent zur Verfügung hat, die sie auf die ihr vorgegebenen Antwortmöglichkeiten verteilen kann. Die Prozente sollen dabei anzeigen, für wie wahrscheinlich die Testperson die Antwortmöglichkeit für richtig hält.

Die vorliegende Arbeit beschäftigt sich mit dem Multiple-Evaluations Format, welches ein Spezialfall des probabilistischen Messens ist und auf einer logarithmischen Auswertung basiert. Bei dieser kann es zu einem negativen Itemscore kommen, wenn die Testperson zu geringe Prozentsätze auf die richtige Antwortmöglichkeit setzt. Durch das spezielle Responseformat sowie durch die Art des Scorings werden eine Reduktion des Rateeffekts sowie eine Erhöhung der Reliabilität und der Validität eines Tests erwartet.

Bezüglich der Validität eines Tests in diesem Format konnten einige Autoren Ergebnisse finden, die darauf schließen lassen, dass neben dem Wissen einer Person auch Persönlichkeitsmerkmale einen Einfluss darauf haben, wie sie die Prozente auf die ihr vorgegebenen Antwortmöglichkeiten verteilt.

Die vorliegende Untersuchung widmete sich dem Einfluss von Geschlecht und Persönlichkeit auf das Antwortverhalten bei Vorgabe eines Tests im Multiplen-Evaluations Format. Es wurde untersucht, ob es neben dem Wissen bzw. der Fähigkeit einer Person noch andere Einflüsse darauf gibt, wie sicher sich eine Person bezüglich ihres Wissens ist und wie realistisch sie dieses einschätzt. Hierfür wurden die Antwortsicherheit (durchschnittliche Itemsicherheit nach Hansen, 1971) sowie das Ausmaß des Profits, den die Testpersonen von einer logarithmischen Auswertung mit Strafpunkten haben, erhoben.

Die Testung erfolgte an einem Wiener Gymnasium an insgesamt 54 Schülern und 108 Schülerinnen der 10ten, 11ten und 12ten Schulstufe. Das Alter der Schüler/innen lag zwischen 15 und 19 Jahren. Aus ökonomischen Gründen wurde die Testung im Papier-

Bleistift Format durchgeführt. Als Tests zur Messung der verbalen Fähigkeiten wurde der Untertest Satzergänzung aus dem I-S-T 2000R im Multiplen-Evaluations Format verwendet. Die Instruktion zum Multiplen-Evaluations Antwortformat erfolgte persönlich und interaktiv. Anhand von drei Übungsfragen wurde die hinter dem Antwortformat stehende Auswertung erklärt. Zur Erfassung der Persönlichkeitsvariablen diente das Hamburger Persönlichkeitsinventar.

Die Hypothesenprüfung zu den Persönlichkeitseinflüssen erfolgte mittels Rangkorrelationen. Auf Grund der vorliegenden geringen Zusammenhänge mit den Persönlichkeitsvariablen Risikobereitschaft und Gewissenhaftigkeit sowie den nicht vorhandenen Zusammenhängen mit den anderen Persönlichkeitsvariablen kann diese Arbeit nicht bestätigen, dass es für die Praxis von Testungen im Multiple-Evaluations Format relevante Einflüsse der Persönlichkeit auf das Antwortverhalten gibt.

Die Geschlechtsunterschiede wurden mittels Welch-Test analysiert. Es konnten bedeutsame Unterschiede zwischen Männern und Frauen in ihrem Antwortverhalten gefunden werden. Es zeigte sich, dass Männer signifikant antwortsicherer sind als Frauen. Dies kann auf die generelle Tendenz von Frauen zurückgeführt werden, ihre Leistung weniger sicher einzuschätzen als Männer. Außerdem profitieren Frauen signifikant mehr von einer logarithmischen Auswertung relativ zu einer linearen, was darauf schließen lässt, dass sie ihr Wissen realistischer einschätzen und wiedergeben.

Der Zusammenhang zwischen der durchschnittlichen Itemsicherheit der Testpersonen und dem Ausmaß ihres Profits von eine logarithmischen Auswertung wurde ebenfalls mittels einer Rangkorrelation berechnet. Der gefundene hohe negative Zusammenhang lässt darauf schließen, dass sich die Antwortsicherheit der Testpersonen nur zu einem geringen Teil auf deren Wissen gründet.

In weiteren Studien erscheint es sinnvoll, den Testpersonen einen mehrmaligen Kontakt mit dem neuen Antwortformat zu ermöglichen sowie die Variable Antwortsicherheit anhand von nicht beantwortbaren Nonsensitems zu erfassen, um die Wissenskomponente zu kontrollieren. Außerdem wäre es von Interesse, die Ergebnisse auch in anderen Teilpopulationen, eventuell auch in realen Testsituationen, zu überprüfen.

Generell bieten sich ausreichende Übungsphasen und ein direktes der Testperson angepasstes Feedback an, um Persönlichkeits- und Geschlechtseinflüsse auf einen Test im Multiplen-Evaluations Format zu reduzieren.

9. Literaturverzeichnis

Andresen, B. (2002). *Hamburger Persönlichkeitsinventar (HPI). Manual.* Göttingen: Hogrefe.

Dirkzwager, A. (1996). Testing with personal probabilities. 11 year-olds can correctly estimate their personal probabilities. *Educational and Psychological Measurement, 56*(6), 957-971.

Dirkzwager, A. (2003). Multiple Evaluation: A new testing paradigm that exorcizes guessing. *International Journal of Testing, 3*(4), 333-352.

Echternacht, G.J., (1972). Use of confidence testing in objective tests. *Review of educational research, 42* (2), 217-236.

Echternacht, G.J., Boldt, R.F. & Sellman, W.S. (1972). Personality influences on confidence test scores. *Journal of Educational Measurement, 9*(3), 235-241.

Ehrlinger, J. & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84,* 5-17.

Hambleton, R.K., Roberts, D.M., Traub, R.E.(1970). A comparsion of the reliability and validity of two methods for assessing partial knowledge in a multiple choice test. *Journal of educational measurement 7*(2), 75-82.

Hannover, B. & Bettge, S.h. (1993). *Mädchen und Technik*. Göttingen: Hogrefe.

Hansen, R. (1971). The Influence of Variables Other than Knowledge on Probabilistic Tests. *Journal of Educational Measurement, 8*(1), 9-14.

Heatherington. I., Daubman, K.A., Bates, C., Ahn, A., Brown, H. & Preston, C. (1993). Two investigations of female modesty in achievement situations. *Sex Roles. 29,739-754.*

Holmes, P. (2002). *Multiple Evaluation versus Multiple Choice as Testing Paradigm-Feasibility, Reliability and Validity in practice. (www.ub.utwente.nl/webdocs/to/1/t0000017.pdf)* (Verfügbar am 15.3.2011)

Koehler, R.A. (1974). Overconfidence on Probabilistic Tests. Journal of Educational Measurement, 11(2), 101-108.

Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* (2., überarb. und erweiterte Auflage). Göttingen: Hogrefe.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007) *Intelligenz-Struktur-Test 2000 R: IST 2000 R.* Göttingen: Hogrefe.

Lienert G.A., Raatz, U. (1998). *Testaufbau und Testanalyse.* (6.Aufl.) Weinheim: Psychologie Verlags Union.

Meijer, R. R. (2003). Diagnosing Item Score Patterns on a Test Using Item Response Theory-Based Person-Fit Statistics. *Psychological Methods*, 8 (1), 72 – 87.

Mondak, J. (2001). Developing valid knowledge scales. *American Journal of Political Science, 45*(1), 224-238.

Moreno, R., Martínez, R. J. & Muñiz, J. (2006). New Guidelines for Developing Multiple-Choice Items. *Methodology*, *2*, 65-72.

Nadeau, R. & Niemi, R.G. (1995). Educated guesses: The process of answering factual knowledge questions in surveys. *Public Opinion Quarterly*, *59*(3), 323-346.

Rasch, D., Kubinger, K.D. & Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Stat Papers, 52*, 219-231.

Rippey, R.M. (1970). A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement, 7*(3), 165-170.

Schaefer, R.E. (1976). Eine Alternative zur konventionellen Methode der Beantwortung und Auswertung von Tests mit Mehrfachantworten. *Diagnostica, 22*, 49-63.

Schaefer, P.S., Williams, C.C., Goodie A.S & Campebell, W.K. (1994). Overconfidence and the big five. *Journal of research in personality, 38*(5), 473-480.

Shuford, E.H. Jr., Albert, A. & Massengill, H.E. (1966*). Admissible Probability Measurement Procedures Psychometrika. 31*(2), 125-145.

Sieber, J.E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. *Journal of personality and social psychology, 30*, 688-694.

Slakter, M.J. (1968). The penalty for not guessing. *Journal of educational measurement, 5*, 141-144.

Teubenbacher, N. (2009). Korreliert ein höherer Rateeffekt mit einem niedrigen Person-fit Index? Unveröffentlichte Diplomarbeit, Universität Wien.

Zimbardo, P.G. & Gerrig, R.J. (2008). *Psychologie* (18.Aufl.). Berlin: Springer.

10. Tabellenverzeichnis

Tabelle 1: Toleranzparameter (Holmes, 2002, S.44) 18

Tabelle 2:Vergleich der Score-Möglichkeiten ME_1, ME_2 und der linearen Auswertung 18

Tabelle 3: Rangkorrelationen des C_T-Werts mit den Persönlichkeitsvariablen des HPIs 36

Tabelle 4: Mittelwertvergleiche von C_T zwischen Männern und Frauen (Welch-Test)..... 37

Tabelle 5: Rangkorrelationen des Ausmaßes des Profits von ME_1 und ME_3 mit den Persönlichkeitsvariablen des HPIs 37

Tabelle 6: Mittelwertvergleiche des „Profits“ von ME_1 und ME_3 zwischen Männern und Frauen (Welch-Test) 38

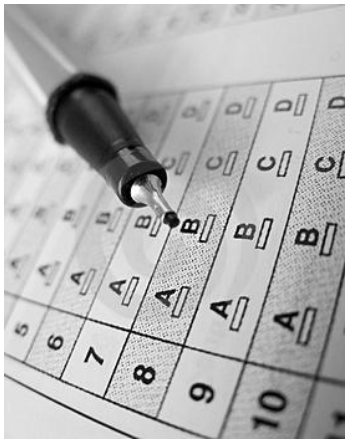
Tabelle 7.: Rangkorrelationen der durchschnittlichen Itemsicherheit mit dem Ausmaß des Profits von einer logarithmischen Auswertung 39

11. Anhang

11.1 Instruktion

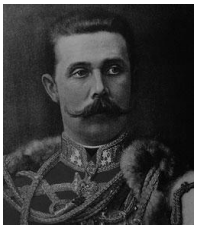
MULTIPLE -EVALUATION

%



Multiple-Choice

- **Erzherzogs Franz Ferdinands letzte Worte waren?**



- ○ **Mehr Licht**
- ✕ **Es ist gar nichts**
- ○ **Ich bin gescheitert**
- ○ **Welch ein Narr bin ich gewesen**
- ○ **Undank ist der Welten Lohn**

Punktevergabe: Multiple Choice

- 1 Punkt pro richtiger Antwort
- 0 Punkte bei falscher Antwort
- Keine Minus Punkte

Multiple-Evaluation

- Man muss sich nicht mehr für eine Antwortmöglichkeit entscheiden
- 100% auf alle Antwortmöglichkeiten verteilen
- Prozente sollen anzeigen, wie sicher man sich ist, dass die Antwortmöglichkeit die richtige ist→ soll wahres Wissen wiederspiegeln

Multiple-Evaluation

Erzherzogs Franz Ferdinands letzte Worte waren?

A: Mehr Licht 0%

B: Es ist gar nichts 50%

C: Ich bin gescheitert 25%

D: Welch ein Narr bin ich gewesen 15%

E: Undank ist der Welten Lohn 10%

100%

Übungsbeispiel 1

- **Wie heißt die Hauptstadt von Österreich?**

A: Graz 0%

B: Wien 100%

C: Bregenz 0%

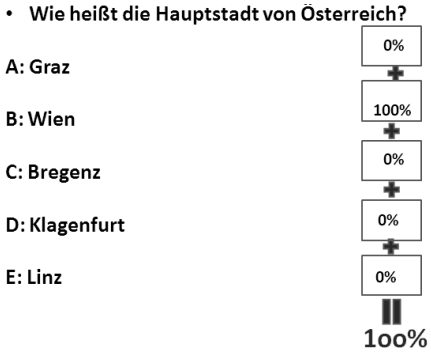
D: Klagenfurt 0%

E: Linz 0%

1oo%

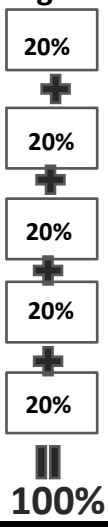
Punktevergabe: Multiple Evaluation

– 100% auf der richtigen Antwortmöglichkeit→ volle Punkteanzahl



Übungsbeispiel 2

- **Welches ist das größte iranische Gebirge?**
- **A: Elburs**
- **B: Elbrus**
- **C: Erebus**
- **D: Ararat**
- **E: Zagros**



Punktevergabe: Multiple Evaluation

– Genau 20% auf die richtige Antwortmöglichkeit = ich weiß die Antwort nicht → **O Punkte**

- **Welches ist das größte iranisches Gebirge?**
- **A: Elburs** 20%
- **B: Elbrus** 20%
- **C: Erebus** 20%
- **D: Ararat** 20%
- **E: Zagros** 20%

100%

Übungsbeispiel 3a

Wie heißt die Hauptstadt von Australien?

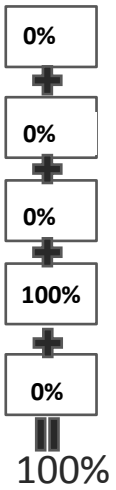
- **A: Sydney** 33%
- **B: Brisbane** 0%
- **C: Canberra** 34%
- **D: Melbourne** 33%
- **E: Perth** 0%

100%

Übungsbeispiel 3b

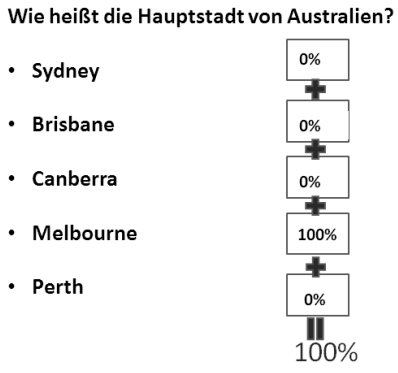
Wie heißt die Hauptstadt von Australien?

- **A: Sydney**
- **B: Brisbane**
- **C: Canberra**
- **D: Melbourne**
- **E: Perth**



Punktevergabe: Multiple Evaluation

– Ab weniger als 20% auf die richtige Antwort→ **Minuspunkte (schwer auszubessern)**



Zusammenfassung

- Ab weniger als 20% bei der richtigen Antwort→ Minuspunkten**(bis zu Minus 3!)**
- Genau 20% → Null Punkte
- Ab mehr als 20% auf die richtigen Antwortmöglichkeit → Plus Punkte

- Ihr könnte die Prozente verteilen wie ihr wollt!
- ABER: Seid bitte möglichst realistisch!

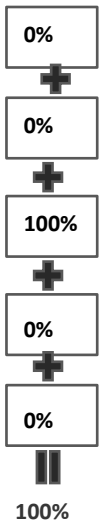
Satzergänzung

- Die folgenden Aufgaben bestehen aus Sätzen, bei denen jeweils ein Wort fehlt. Für jeden der Sätze werden euch fünf Lösungsmöglichkeiten vorgeschlagen.
- Bitte verteilt eure 100 Prozent, wie in den vorherigen Übungsbeispielen besprochen, so dass sie anzeigen wie sicher ihr euch seid, dass die entsprechende Antwortmöglichkeit die Richtige ist.
- **Es ist immer nur eine Antwortmöglichkeit die Richtige!**

BEISPIELE

Ein Kaninchen hat am meisten Ähnlichkeit mit einem (einer)…?

a) Katze
b) Eichhörnchen
c) Hasen
d) Fuchs
e) Igel



- **Das Gegenteil von Hoffnung ist…?**

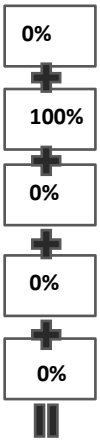
a) Trauer

b)Verzweiflung

c) Elend

d) Liebe

e) Hass



BITTE KONTROLLIERT, DASS SICH EURE PROZENT PRO FRAGE IMMER AUF **100** AUSGEHEN!

11.2 Instruktionsblatt

MULTIPLE-EVALUATION

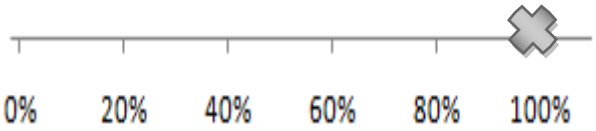
Beim Multiplen-Evaluations Format müsst ihr euch nicht wie beim Multiple-Choice Format für eine Antwortmöglichkeit entscheiden, sondern ihr habt 100% zu Verfügung, die ihr auf alle Antwortmöglichkeiten verteilen könnt. Diese Prozente sollen anzeigen, wie sicher ihr euch seid, dass die jeweilige Antwortmöglichkeit die richtige ist.

PUNKTEVERGABE

Wie heißt die Hauptstadt von Australien? a)Sydney b)Melbourne c)Canberra d)Brisbane e)Perth

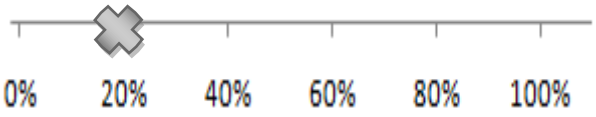
- Wenn ihr euch ganz sicher seid, setzt 100% auf die richtige Antwortmöglichkeit →**volle Punkteanzahl**

c) Canberra



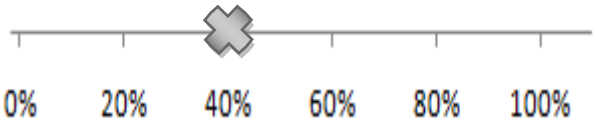
- Wenn ihr keine der Antwortmöglichkeiten ausschließen könnt, verteilt die Wahrscheinlichkeiten gleich (a-e je 20%)→ 20% auf der richtigen Antwortmöglichkeit = „Ich weiß die Antwort nicht“ → **0 Punkte**

a)Sydney
b)Melbourne
c)Canberra je
d)Brisbane
e)Perth



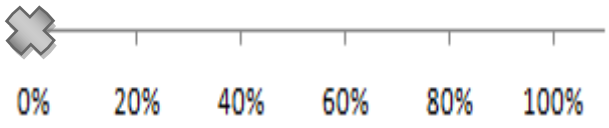
- Bei über 20% auf der richtigen Antwortmöglichkeit beginnt ihr **Pluspunkte** zu bekommen.

c)Canberra



- Wenn ihr weniger als 20% auf die richtige Antwortmöglichkeit setzt → **Minuspunkte** (bei 0% auf der richtigen Antwortmöglichkeit maximale Anzahl an Minuspunkten)

c)Canberra



WICHTIG: **Eure vergebenen Prozente müssen sich immer auf <u>100</u> ausgehen!**

Lebenslauf

Angaben zur Person

Name	Livia Jenner
E-Mail	liviaj@gmx.at
Geburtsort	Wien
Geburtsdatum	09.10.1985

Berufserfahrung

Seit Juli 2011	Caritas Socialis – Betreuung zu Hause
	Einsatzkoordinatorin
Mai 2010 – Juli 2011	Wifi Wien – Seminarbetreuung
	geringfügige Beschäftigung, Trainer/innen Betreuung
Nov. 2009 - Mai 2010	Praktikum Geriatriezentrum Donaustadt
	Psychologisch Diagnostische Testung, Psychologische Betreuung und Begleitung
Mai 2006 – Aug. 2008	Denzel – Denzeldrive Carsharing
	geringfügige Beschäftigung als Callcenter Agent, Kundenbetreuung, administrative Tätigkeiten

Ausbildung

Seit Okt. 2004	Universität Wien Diplomstudium der Psychologie
Sept. 2000- Juni 2004	ORG Hegelgasse 14
Sept.1996 – Juni 2000	BRG Radetzky Straße