Wissensdiagnostik mit Multiple-Choice und Multipler Evaluation - ein Vergleich logarithmischer und linearer Auswertefunktionen

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Erika Heidi Enders

aus Hilden

Düsseldorf, September 2011

Aus dem Institut für Experimentelle Psychologie
Abteilung für Diagnostik und Differentielle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Jochen Musch
Koreferentin: Prof. Dr. Ute J. Bayen, PhD.

Tag der mündlichen Prüfung: 11. November 2011

Danksagung

An erste Stelle danke ich den vielen tausend Internetnutzern, die durch ihre Teilnahme an den Experimenten die Erhebung der Daten ermöglicht haben. Allen Personen, die die Testapplikation getestet haben und uns wertvolle Hinweise zur Verbesserung lieferten, danke ich ebenfalls sehr. Bei Herrn Andreas Fels (APIX Internet Services GmbH, Köln) bedanke ich mich für die Bereitstellung des Webservers mit Datenbank sowie für die Schaltung von Werbebannern auf diversen Webseiten von APIX-Projekten.

Meinem Betreuer Herrn Prof. Jochen Musch danke ich sehr herzlich für seine uneingeschränkte Unterstützung und Bereitstellung von Hilfsmitteln. Ich verdanke ihm zudem wertvolle wissenschaftliche Anregungen. Bei Frau Prof. Ute Bayen bedanke ich mich sehr für die Übernahme der Zweitbegutachtung.

Herr Reinald Witsch verdient Dank für seine Beratung bei der Auswahl der Items. Bei Herrn Prof. Michael Schindler und Herrn Dr. Graham Holmwood bedanke ich mich für das Korrekturlesen der Summary. Meiner Mutter bin ich dankbar für ihr Korrekturlesen und ihre beständige motivierende Unterstützung.

Mein ganz besonderer Dank gilt jedoch Herrn Ingo Kühlborn. Ich danke ihm für die ausgezeichnete Programmierung der Webapplikation sowie dem geduldigen, ausgesprochen konstruktiven Korrekturlesen meiner Entwürfe. Darüberhinaus danke ich ihm für seine stete Motivation, mit deren Hilfe ich auch Tiefpunkte überstehen konnte.

Zusammenfassung

Das Multiple-Choice-Verfahren ermöglicht eine objektive Messung von Wissen und eine ökonomische Auswertung. Insbesondere bei schwierigen Items können jedoch Rateprozesse die Varianz in den beobachtbaren Testwerten erhöhen, und wichtige diagnostische Information geht verloren, wenn Teilwissen nicht erfasst wird. Bei der Multiplen Evaluation wird das Wissen der Testteilnehmer differenzierter erfasst, indem ihre Antwortsicherheit bezüglich sämtlicher Antwortoptionen erfragt wird (Dirkzwager, 2003). Dabei kann durch den Einsatz von logarithmischen Auswertefunktionen und Strafzahlungen sichergestellt werden, dass ein Teilnehmer sein Testergebnis nur maximieren kann, wenn er seine Antwortsicherheit unverfälscht berichtet (Shuford, Albert & Massengill, 1966). In drei Experimenten wurde die Frage untersucht, ob das Verfahren der Multiplen Evaluation eine bessere Wissensdiagnostik ermöglicht als das Multiple-Choice-Verfahren.

Im ersten Experiment zeigte sich, dass die Multiple Evaluation zu einer gegenüber dem Multiple-Choice-Verfahren erhöhten Reliabilität führte. Dies war auch dann noch der Fall, wenn dem Testteilnehmer die logarithmische Auszahlungsfunktion lediglich vorab kommuniziert wurde, ohne dass eine itemspezifische Rückmeldung über die Punktauszahlungen erfolgte. Darüber hinaus zeigte sich, dass die erzielte Reliabilitätsverbesserung nicht an die zwar differenziertere, dafür aber auch zeitintensivere Erfassung der Antwortsicherheit mithilfe von Schiebereglern geknüpft war. Vielmehr verbesserte sich die Reliabilität bereits bei einer Erfassung der Antwortsicherheit mithilfe eines einzelnen Mausklicks. Dazu wurde ein Antwortdreieck verwendet, welches eine Auswahl aus einer diskreten Menge von 16 Antwortkategorien zur simultanen Erfassung der Antwortsicherheit in drei zur Verfügung stehenden Antwortoptionen ermöglichte. Das Ergebnismuster legt nahe, dass ein Reliabilitätszugewinn mithilfe der Multiplen Evaluation nicht auf computergestützte Testungen beschränkt sein muss.

Im zweiten Experiment zeigte sich, dass logarithmische Auswertefunktionen mit zu hohen Strafzahlungen sowohl die Reliabilität als auch die Validität von Tests mit Multipler Evaluation beeinträchtigen. Die höchste Reliabilität und Validität wurde mit der Multiplen Evaluation unter Verwendung einer linearen Auswertefunktion beobachtet, die gar keine Strafpunkte vorsah.

Im dritten Experiment wurde der Einfluss der Anzahl der Antwortoptionen auf die Reliabilität und die Validität der konkurrierenden Antwortprozeduren untersucht. Bei beiden Verfahren verschlechterte sich die Reliabilität, nicht jedoch die Validität mit kleiner werdender Zahl der Antwortoptionen.

In allen Experimenten wurde die Güte der Kalibrierung der Teilnehmer mithilfe eines Realismusindex beurteilt (Holmes, 2002). Durch eine nachträgliche Korrektur der Antwortsicherheiten auf der Basis dieses Index konnte die Reliabilität signifikant verbessert werden. Eine Verbesserung auch der Validität zeigte sich, wenn die Auswertung mithilfe einer logarithmischen Funktion erfolgte, die hohe Strafzahlungen vorsah.

Zusammenfassend belegen die Ergebnisse, dass die Qualität der Wissensdiagnostik durch den Einsatz der Multiplen Evaluation vor allem bei schwierigen Items und unter Verwendung von linearen Auswertefunktionen oder logarithmischen mit nicht zu hohen Strafzahlungen verbessert werden kann.

Summary

Multiple choice techniques permit an objective measurement of knowledge and an economic scoring. Especially for difficult items guessing procedures can, however, increase the variance in the observed test values, and important diagnostic information gets lost when partial knowledge is not captured. With multiple evaluation the knowledge of a participant is captured more precisely by asking for his percentage confidence for each possible answer (Dirkzwager, 2003). Through the use of logarithmic scoring functions and penalty payments it can be ensured that a participant can only maximize his test result if he makes an unbiased demonstration of his confidence percentages (Shuford, Albert & Massengill, 1966). In three experiments it was investigated whether multiple evaluation offers a better assessment of knowledge than multiple choice.

The first experiment showed that the multiple evaluation procedure led to a higher reliability than multiple choice. This was also true when a logarithmic scoring function was used without providing feedback about the item specific scores and penalties beforehand. In addition it was observed that an increase in reliability was not linked to the differentiated, though time-consuming, capturing of confidences by means of sliders. Indeed, the reliability already increased when capturing confidences by a single mouse click. For that purpose an answer triangle was used, which offered a discrete number of 16 answer categories for simultaneously capturing confidences in all three possible answers. The resulting pattern suggests that increasing the reliability using multiple evaluation is not necessarily limited to a computer based assessment.

The second experiment showed that exceedingly high penalty payments in logarithmic scoring functions reduce both the reliability and the validity of a test with multiple evaluation. The highest reliability and validity were observed with a linear scoring function that does not provide for any penalty payments.

In the third experiment the influence of the number of possible answers on the reliability and validity of both competing answer techniques was investigated. With a decreasing number of possible answers in both procedures the reliability decreased whereas the validity was not affected.

In all experiments the participants′ level of calibration was judged by an individual realism index (Brown & Shuford, 1973; Holmes, 2002). The reliability was improved significantly by an a posteriori correction of the confidences based on this index. The validity was improved, too, if the scoring used a logarithmic function with high penalty payments.

In conclusion, the results prove that the quality of the assessment of knowledge especially for difficult items can be improved by using multiple evaluation together with linear or logarithmic scoring functions with moderate penalty payments.

Inhaltsverzeichnis

1 Einleitung

Zur Messung von Wissen wird häufig das Antwortwahlverfahren (Multiple-Choice) herangezogen. In einem solchen Test beantwortet ein Teilnehmer eine Frage, indem er die aus seiner Sicht am wahrscheinlichsten richtige Antwort aus mehreren Optionen auswählt. Da eine Antwort nur entweder richtig oder falsch sein kann, ist das Verfahren einfach auszuwerten. Aus diesem Grund wird es auch gerne in Quizsendungen eingesetzt. Gerade ein Wissensquiz mit Multiple-Choice veranschaulicht die Schwächen des Verfahrens. Ein Kandidat hat eine gute Chance, wenn er die richtige Antwort nicht sicher weiß, diese zu erraten. Verfügt er dabei über Teilwissen, durch das er Antwortoptionen mit hoher Wahrscheinlichkeit als falsch ausschließen kann, erhöht sich seine Chance, beim Raten erfolgreich zu sein. Der Gewinn eines Kandidaten in einem Quiz mit Multiple-Choice wird also nicht nur durch sein Wissen bestimmt, sondern auch durch sein Rateglück.

Das Ziel der Messung von Wissen in akademischen Bereichen ist es, den wahren Wissensstand eines Testteilnehmers zu erfassen. Obwohl die Schwächen des Verfahrens bekannt sind, wird hierzu dennoch häufig Multiple-Choice eingesetzt, da es eine objektive und ökonomische Messung ermöglicht. Jedoch wird meist darüber hinweg gesehen, dass das Raten der Teilnehmer die Varianz der beobachtbaren Testwerte eines Tests erhöht und wichtige diagnostische Information verloren geht, da Teilwissen nicht gemessen werden kann (Koele, De Boo & Verschure, 1987; Shuford & Massengill, 1966). Viele Forscher suchten nach Möglichkeiten, diese Schwächen von Multiple-Choice zu beheben, bisher jedoch ohne zufriedenstellende Ergebnisse (Abedi & Bruno, 1989). Schon in der Mitte des 20. Jahrhunderts wurde das konkurrierende Antwortbewertungsverfahren (Multiple Evaluation) vorgeschlagen (De Finetti, 1965; Shuford & Brown, 1975). Anders als bei einem Test mit Multiple-Choice wählt ein Teilnehmer bei einem Test mit Multipler Evaluation nicht nur die Antwortoption, die er mit der höchsten Wahrscheinlichkeit als die richtige ansieht, sondern er gibt seine Antwortsicherheit für alle vorgegebenen Antwortoptionen eines Items an (Dirkzwager, 2003). Mit diesem Antwortverfahren kann ein Teilnehmer Teilwissen angeben und auch Nichtwissen einräumen. Die vorliegende Arbeit hatte zum Ziel, die Frage zu beantworten, ob das Antwortbewertungsverfahren (Multiple Evaluation) deshalb als geeigneter

eingestuft werden kann, das Wissen eines Teilnehmers zu quantifizieren, als Multiple-Choice.

Die Auszahlung von Punkten anhand von speziellen logarithmischen Auswertefunktionen stellt bei Tests mit Multipler Evaluation sicher, dass ein Teilnehmer nur dann seine maximale Gesamtpunktsumme erreicht, wenn er bei Teil- oder Nichtwissen nicht rät, sondern seine Antwortsicherheit unverfälscht angibt (Shuford, Albert & Massengill, 1966). Dies wird erreicht durch den Abzug von Punkten, wenn ein Testteilnehmer nur eine geringe Antwortsicherheit in die richtige Antwortoption angibt. Für einen Teilnehmer ist es unter einer solchen Auswertung die einzige Strategie, seine Antwortsicherheiten vollkommen unverfälscht zu berichten, wenn er sein Testergebnis maximieren will. Das herkömmliche Multiple-Choice-Verfahren beruht nicht auf einer solchen Auswertefunktion, die die unverfälschte Reproduktion des Wissens belohnt. Die erwartete Gesamtpunktsumme ist vielmehr dann am höchsten, wenn sich der Testteilnehmer stets für die ihm am plausibelsten erscheinende Antwortoption entscheidet. Das Gleiche gilt, wenn bei einem Test mit Multipler Evaluation eine lineare Auswertefunktion verwendet wird. In diesem Fall besteht für den Testteilnehmer kein Anreiz, seine Antwortsicherheiten unverfälscht zu berichten, d.h. auch Teil- oder Nichtwissen einzuräumen. Es war daher anzunehmen, dass die Varianz der beobachtbaren Testwerte, die durch eine verfälschte Wiedergabe von Antwortsicherheiten entsteht, bei einem Test mit Multipler Evaluation mit einer logarithmischen Auswertung im Vergleich zu einer linearen reduziert werden kann. Daher sollte mit dem Antwortverfahren Multiple Evaluation und einer logarithmischen Auswertung die psychometrische Qualität eines Tests im Vergleich zu einer linearen Auswertung verbessert werden können (Dirkzwager, 2003; Shuford, Albert & Massengill, 1966; Shuford & Brown, 1975). Für die vorliegende Arbeit war es eine zentrale Frage, ob die Auswertung mithilfe einer logarithmischen Funktion tatsächlich zu einer Verbesserung der Güte eines Tests führt. Denkbar wäre es nämlich auch, dass die Teilnehmer bei einer Auswertung mit einer logarithmischen Funktion häufig Strafpunkte erhalten. Dies wäre dann zu erwarten, wenn die Teilnehmer aufgrund einer nicht perfekten Kalibrierung häufig nur geringe Antwortsicherheiten in die richtige Antwortoption wiedergäben. Diese Strafzahlungen könnten ihrerseits die Varianz der beobachtbaren Testwerte erhöhen und die psychometrische Güte eines Tests reduzieren. Unter diesen Umständen wäre

es deshalb möglich, dass eine gewünschte Verbesserung der Güte eines Tests empirisch auch bei der Verwendung einer linearen Auswertefunktion beobachtet werden kann.

Ausschlaggebend für den effektiven Einsatz einer logarithmischen Auswertefunktion scheint demnach die möglichst perfekte Kalibrierung eines Teilnehmers zu sein. Bei der Wissensdiagnostik mit Multipler Evaluation kann die Güte der Kalibrierung anhand eines individuellen Realismusindex (Brown & Shuford, 1973; Holmes, 2002) bestimmt werden. Bei einer Korrektur auf der Basis dieses Index werden die von einem Teilnehmer berichteten Antwortsicherheiten seiner tatsächlichen subjektiven Antwortsicherheit angenähert. Diese Korrektur wurde vorgeschlagen, um die Varianz der beobachtbaren Testwerte eines Tests, die durch eine verfälschte Wiedergabe der Antwortsicherheit hervorgerufen wurde, nachträglich zu reduzieren (Brown & Shuford, 1973; Holmes, 2002). Die vorliegende Arbeit hatte das Ziel, zu klären, ob diese Korrektur der Antwortsicherheiten auf der Basis des individuellen Realismusindex deshalb ein geeignetes Verfahren ist, um die Reliabilität und die Validität eines Tests zu verbessern.

Bis vor wenigen Jahren war es ein entscheidender Nachteil der Multiplen Evaluation, dass dieses Antwortverfahren ohne Einsatz eines Computers nur schwierig zu realisieren ist und Computer früher nicht in ausreichendem Maße zur Verfügung standen (Rippey & Voytovich, 1983). Dieses Problem besteht heute aufgrund der Fortschritte und der Verbreitung der Computertechnik nicht mehr, sodass eine systematische Überprüfung des Antwortverfahrens Multiple Evaluation vorgenommen werden kann. In der vorliegenden Arbeit erfolgte die Durchführung der Experimente computergestützt in Form von Webexperimenten.

Das folgende zweite Kapitel geht genauer auf die Schwachpunkte von Multiple-Choice ein. Das dritte Kapitel beschreibt Ansätze, das Antwortverfahren Multiple-Choice zu verbessern. Das alternative Antwortverfahren Multiple Evaluation wird im vierten Kapitel vorgestellt. Das sich daran anschließende fünfte Kapitel gibt einen Überblick über den Stand der Forschung zur Multiplen Evaluation. Die Ziele der Experimente dieser Arbeit werden in Kapitel sechs dargelegt. In Kapitel sieben werden Vor- und Nachteile von Webexperimenten diskutiert. Das in den Experimenten verwendete Testmaterial wird in Kapitel acht beschrieben. Die drei Kapitel neun bis elf enthalten die Methodik, die Ergebnisse

sowie die Diskussion der Ergebnisse der drei im Rahmen dieser Arbeit durchgeführten Experimente. Eine abschließende Diskussion erfolgt in Kapitel zwölf. Das Kapitel 13 gibt einen Ausblick auf weitere zu stellende Forschungsfragen. Zur besseren Lesbarkeit wurde auf eine geschlechtsspezifische Differenzierung verzichtet. Mit den Personenbezeichnungen sind immer sowohl Frauen als auch Männer gemeint.

2 Schwächen einer Wissensdiagnostik mit Multiple-Choice

In der Mitte des 19. Jahrhunderts begannen Prüfer, mündliche Examen durch schriftliche zu ersetzen (Bradbard, Parker & Stone, 2004). Ein Antwortverfahren, das seitdem eine große Bedeutung erlangt hat, ist Multiple-Choice. Ein Multiple-Choice-Test besteht aus Mehrfachwahlaufgaben. Bei diesem Aufgabentyp soll ein Testteilnehmer aus einem Angebot von mehreren Antwortoptionen, das aus einer einzigen richtigen Antwort und einer Anzahl von Distraktoren besteht, die richtige Antwort auswählen (Dorsch, Häcker & Stapf, 1992). Multiple-Choice ist auch heute noch ein vielfach eingesetztes Verfahren, besonders aufgrund seiner guten Objektivität und Ökonomie. Das Antwortverfahren wurde aber gleichzeitig auch häufig kritisiert. Das Ausmaß dieser Kritik verdeutlicht beispielsweise Wood (1991, S. 32): "*No assessment technique has been rubbished quite like multiple choice, unless it be graphology.*" Ein Hauptgrund für diese scharfe Kritik ist, dass das dichotome Antwortverfahren Teilnehmer dazu auffordert zu raten, wenn sie die richtige Antwort nicht sicher wissen. Die Varianz der beobachtbaren Testwerte, die durch das Raten der Teilnehmer entsteht, kann jedoch nicht identifiziert werden. Darüber hinaus kann Teilwissen nicht quantifiziert werden, wodurch wichtige diagnostische Information verloren geht. Im Folgenden wird auf diese beiden Kritikpunkte genauer eingegangen.

Wenn ein Teilnehmer ein Item in einem Multiple-Choice-Test richtig löst, bedeutet das nicht unbedingt, dass er die richtige Antwort auch tatsächlich gewusst hat. Er könnte sie auch nur erraten haben, denn die Wahrscheinlichkeit, ein Item nur durch Raten richtig zu beantworten, beträgt 1/k, wobei k für die Anzahl der Antwortoptionen steht. Bei einem Item mit drei Antwortoptionen beträgt diese

Wahrscheinlichkeit also gerundet 33%. Bei einem Test, der aus beispielsweise 36 Items mit je drei Antwortoptionen besteht, beträgt die zu erwartende Punktsumme eines Teilnehmers ohne jegliches Wissen zwölf Punkte, wenn er für jede richtige Antwort einen Punkt erhält und für jede falsche null Punkte. Verfügt ein Teilnehmer zudem über Teilwissen, durch das er Antwortoptionen als sicher falsch ausschließen kann, so kann er sein Ergebnis noch weiter verbessern. Denn durch einen Ausschluss von Antwortoptionen erhöht er die Wahrscheinlichkeit, die richtige Antwortoption zu erraten (Coombs & Womer, 1956). Der Versuch, bei Nicht- oder Teilwissen die richtige Antwortoption zu erraten, ist also für einen Teilnehmer eine logische und angemessene Teststrategie. Ein Testleiter hingegen, der die Aufgabe hat, die Teilnehmer aufgrund ihres Wissens zu differenzieren, steht vor einem unlösbaren Problem. Sein Ziel, den wahren Wissensstand eines Testteilnehmers zu messen, kann er nicht erreichen, da er nicht unterscheiden kann, ob eine richtige Antwort gewusst wurde oder nur Rateglück war (Budescu & Bar-Hillel, 1993). Die Varianz der beobachtbaren Testwerte, die durch Rateprozesse erzeugt wird, ist in einem Multiple-Choice-Test zudem nicht bei allen Teilnehmern gleich. Ob und wie viel ein Teilnehmer rät, ist individuell verschieden und abhängig von verschiedenen Faktoren. Je geringer beispielsweise die Fähigkeit eines Teilnehmers ist, je schwieriger die Items sind und je kürzer die Zeit ist, die ihm zur Bearbeitung des Tests zur Verfügung steht, desto häufiger wird er raten (Bortz & Döring, 2006; Wood, 1991). Damit stellen Rateprozesse eine unsystematische Störvarianz dar und dürfen nicht vernachlässigt werden. Viele Forscher, beispielsweise Shuford und Massengill (1966), kritisierten, dass die durch Rateprozesse hervorgerufene Varianz der beobachtbaren Testwerte in einem Multiple-Choice-Test häufig ignoriert würde. Für die Wissensdiagnostik wäre es aber wichtig, diese Varianz zu reduzieren. Sicher möchte kein Patient von einem Arzt behandelt werden, der seine Diagnose unter verschiedenen Möglichkeiten nur erraten hat, sondern von einem, der die Diagnose aufgrund der Krankheitssymptome sicher gestellt hat.

Nach Ben-Simon, Budescu und Nevo (1997) kann ein Testteilnehmer sich typischerweise in einem von drei Wissensstadien befinden:

1. Er kennt die Antwort und ist sich deren völlig sicher. Er besitzt Vollwissen.

2. Er weiß nur Teile der Antwort oder ist unsicher bezüglich der Antwort. Er verfügt über Teilwissen.
3. Er besitzt kein Wissen über die richtige Antwort. Sein Wissensstand ist Nichtwissen.

In einem herkömmlichen Multiple-Choice-Verfahren kann ein Testteilnehmer nur das Stadium Vollwissen angeben. Häufig verfügt ein Teilnehmer aber nur über Teilwissen. Dieses kann er in einem Multiple-Choice-Test jedoch nicht wiedergeben. Auch die Möglichkeit, Nichtwissen einzuräumen, besteht in der Regel nicht. In einem herkömmlichen Multiple-Choice-Test geht damit diagnostische Information verloren, deren Quantifizierung aber sowohl für Lehrende als auch für Lernende wichtig und aufschlussreich wäre (De Finetti, 1965).

3 Ansätze zur Verbesserung von Multiple-Choice

Viele Forscher verfolgten bereits das Ziel, das Antwortverfahren Multiple-Choice zu verbessern. Sie modifizierten dazu Items, Instruktionen und Auswertungsregeln (Ben-Simon, Budescu & Nevo, 1997). Im Folgenden werden einige dieser Ansätze exemplarisch vorgestellt.

Um die Varianz der beobachtbaren Testwerte aufgrund des Ratefehlers zu reduzieren, kann die Anzahl der Antwortoptionen eines Items erhöht und so die Wahrscheinlichkeit, die richtige Antwort zu erraten, verringert werden (Bruno & Dirkzwager, 1995). Bei einem Multiple-Choice-Test werden deshalb häufig Items mit vier oder fünf Antwortoptionen verwendet. Eine hohe Anzahl von gleichwertigen Distraktoren zu erstellen ist jedoch eine komplexe Aufgabe (Dressel & Schmid, 1953). Haladyna und Downing (1993) vermuteten, dass drei Antwortoptionen, eine richtige Antwort und zwei Distraktoren, in den meisten Fällen eine natürliche Grenze für die Konstrukteure von Items seien. Studien konnten zudem zeigen, dass eine Erhöhung der Anzahl der Antwortoptionen die psychometrische Qualität eines Tests nicht generell verbessert. Burton (2001) fand beispielsweise, dass bei einem aus 60 Items bestehenden Test die Reliabilität durch eine Erhöhung der Anzahl der Antwortoptionen von vier auf fünf nicht weiter verbessert werden konnte. Grier (1975) zeigte, dass die Reliabilität eines Tests dann maximal war, wenn er Items mit

drei Antwortoptionen einsetzte. Rodriguez (2005) führte eine Metaanalyse über 27 Studien durch. Er betrachtete Itemschwierigkeiten, Trennschärfen, Reliabilitäten und Validitäten und kam ebenfalls zu dem Schluss, dass in den meisten Fällen Multiple-Choice-Tests, die aus Items mit drei Antwortoptionen konstruiert sind, aussagekräftige Testergebnisse liefern. Die Güte eines Tests könne eher dadurch verbessert werden, dass statt Items mit beispielsweise fünf Antwortoptionen, in vergleichbarer Testzeit, eine höhere Anzahl von Items mit nur drei Antwortoptionen verwendet werde. Bruno und Dirkzwager (1995) kamen durch die Betrachtung eines informationstheoretischen Modells ebenfalls zu dem Schluss, dass drei Antwortoptionen optimal seien. Aufgrund der vorliegenden Befunde lässt sich deshalb folgern, dass eine Erhöhung der Anzahl der Distraktoren auf mehr als zwei die psychometrische Qualität eines Multiple-Choice-Tests in der Regel nicht verbessert.

Durch eine Erhöhung der Anzahl der richtigen Antworten eines Items kann ebenfalls die Wahrscheinlichkeit verringert werden, ein Item nur aufgrund von blindem Raten zu lösen. Bei einem Item mit einer richtigen Antwort und vier Distraktoren beträgt die Ratewahrscheinlichkeit 20%. Besteht ein Item dagegen aus zwei richtigen Antworten und drei Distraktoren, so beträgt diese Wahrscheinlichkeit nur noch 10%. Kubinger und Gottschall (2007) konnten empirisch zeigen, dass die Erhöhung der Anzahl der richtigen Antworten von eins auf zwei die Schwierigkeit der Items erhöht. Wie bereits ausgeführt, ist die Erstellung von gleichwertigen Distraktoren jedoch eine schwierige und aufwendige Aufgabe. Items mit zwei richtigen Antwortoptionen zu erstellen, stellt dabei wohl eine noch größere Herausforderung dar.

Bei einem Multiple-Choice-Test kann in der Regel eine Verbesserung der Reliabilität erzielt werden, indem die Anzahl der Items erhöht wird. Durch die Verlängerung eines Tests können aber Störgrößen wie Ermüdungserscheinungen, eine mangelnde Konzentration und eine sinkende Testmotivation der Teilnehmer wiederum zu Messfehlern führen (Rost, 2004). Zudem verlängert sich durch eine Erhöhung der Anzahl der Items auch die Durchführungszeit eines Tests. Die Verlängerung von Tests mit dem Ziel, die Reliabilität zu verbessern, birgt also den Nachteil einer verringerten Ökonomie sowie die Gefahr einer zusätzlichen Erhöhung der Varianz der beobachtbaren Testwerte.

Ein weiterer Ansatz in einem Multiple-Choice-Test, die Varianz der beobachtbaren Testwerte, die durch Rateprozesse erzeugt wird, zu reduzieren, ist eine Ratekorrektur. Einer Ratekorrektur liegt die Annahme zugrunde, dass falsche Antworten nicht durch fehlerhaftes Wissen, sondern nur durch Nichtwissen zustande kommen. Wie viele Items ein Testteilnehmer durch Raten richtig beantworten konnte, ist gegenüber den durch Wissen richtig beantworteten Items nicht zu unterscheiden. Deshalb wird die Anzahl der falsch beantworteten Items herangezogen, um die der richtig erratenen zu kompensieren (Wood, 1991). Eine Ratekorrektur der Punktsummen kann beispielsweise anhand der folgenden Formel durchgeführt werden (Leclerq, 1983):

$$P = R - \frac{F}{k-1}$$

Dabei bedeutet:
P=Punktsumme
R=Anzahl der richtigen Antworten
F=Anzahl der falschen Antworten
k=Anzahl der Antwortoptionen

Bei einem Test, der aus 36 Items mit je drei Antwortoptionen besteht, beträgt die Punktsumme eines Teilnehmers, der 20 Items richtig lösen konnte, vor der Korrektur 20 Punkte. Da er 16 Items nicht richtig gelöst hat, werden ihm aufgrund der Korrektur 16/(3-1), also 8 Punkte abgezogen. Seine Punktsumme beträgt also nach der Korrektur 12 Punkte. Prihoda, Pinckard, McMahan und Jones (2005) konnten eine Verbesserung der Validität eines Tests durch eine Ratekorrektur zeigen. Jedoch kann diese Ratekorrektur Teilnehmer verunsichern und demotivieren, da sie unfair ist. Wenn ein Teilnehmer über Teilwissen verfügt, wird die Ratekorrekturformel unterkorrigieren, da der Teilnehmer beim Raten erfolgreicher ist als beim allein auf Zufall basierten Raten. Im Falle von fehlerhaftem Wissen wird sie überkorrigieren, da in diesem Fall der Teilnehmer nicht nur die Punkte aufgrund seiner falschen Antworten verliert, sondern weitere Punkte aufgrund der Ratekorrektur (Burton, 2001; Leclerq, 1983; Rowley & Traub, 1977). Außerdem muss kritisch gesehen

werden, dass Raten nur kompensiert, aber nicht verhindert wird (Coombs & Womer, 1956).

Coombs und Womer (1956) entwickelten ein sogenanntes Eliminierungsverfahren, das Raten verhindern und darüber hinaus Teilwissen messen sollte. Sie verwendeten Items mit einer richtigen Antwortoption und drei Distraktoren. Die Testteilnehmer beantworteten ein Item, indem sie möglichst alle Distraktoren ankreuzten. Für jeden richtig erkannten Distraktor erhielten sie einen Punkt. Markierten die Teilnehmer fälschlicherweise die richtige Antwort, wurden ihnen drei Punkte abgezogen. Damit betrug die mögliche Punktauszahlung je Item zwischen -3 und 3 Punkten. Dieses Eliminierungsverfahren hat gegenüber dem herkömmlichen Multiple-Choice-Verfahren den Vorteil, dass Raten nicht belohnt wird, denn der Erwartungswert der Punktsumme, die allein durch blindes Raten erreicht werden kann, beträgt null Punkte. Zudem kann ein Testteilnehmer durch den Ausschluss von Distraktoren Teilwissen zeigen. Als Kritik ist jedoch festzuhalten, dass dieses Verfahren Fehler bei der Beantwortung der Items begünstigt, da alle Distraktoren markiert werden müssen (Akeroyd, 1982), und negatives Denken fördert, da die Aufmerksamkeit auf die falschen Antwortoptionen gerichtet wird und nicht auf die richtigen (Wood, 1991). Coombs und Womer (1956) konnten durch den Einsatz ihres Eliminierungsverfahrens eine Verbesserung der Reliabilität eines Tests um 20% gegenüber einem konventionellen Multiple-Choice-Verfahren zeigen. Dabei fanden sie die deutlichste Verbesserung bei schwierigen Items. Bradbard und Green (1986) sowie Bradbard, Parker und Stone (2004) konnten diese Ergebnisse von Coombs und Womer (1956) jedoch nicht reproduzieren.

Zusammenfassend konnte keiner der vorgestellten Ansätze die Varianz der beobachtbaren Testwerte, die durch das Raten der Teilnehmer erzeugt wird, zufriedenstellend reduzieren. Auch Teilwissen konnte mit den meisten Verfahren nicht quantifiziert werden.

4 Wissensdiagnostik mit Multipler Evaluation

Schon in der Mitte des 20. Jahrhunderts wurde das alternative Antwortverfahren Multiple Evaluation mit dem Ziel vorgeschlagen, die Varianz der beobachtbaren Testwerte eines Tests, die durch Rateprozesse hervorgerufen wird, zu reduzieren und auch Teilwissen zu quantifizieren (Dirkzwager, 2003). Im Folgenden wird ausgeführt, wie bei einem Test mit Multipler Evaluation die Messung des Wissens eines Teilnehmers in Form von Antwortsicherheit oder Konfidenzurteilen erfolgt.

4.1 Erhebung von Antwortsicherheit

Mit dem Ziel, die diagnostische Qualität von Wissenstests zu erhöhen, schlug De Finetti (1965) vor, die subjektive Antwortsicherheit eines Testteilnehmers zu erheben. Ein Teilnehmer soll demnach seine Sicherheit, mit der er jede einzelne Antwortoption eines Items für richtig hält, angeben. Wissen wird so operationalisiert als die Fähigkeit eines Testteilnehmers, die Richtigkeit der Antwortoptionen eines Items zu bewerten (Shuford & Brown, 1975). Eine numerische Antwortsicherheit ist daher eine komplexe Funktion des vorhandenen Wissens eines Teilnehmers zu dem Zeitpunkt, an dem diese erfasst wird. Antwortsicherheit ist also mit einem Zeitpunkt verbunden und kann variieren, z.B. durch die Zunahme des vorhandenen Wissens (Shuford & Brown, 1975). Durch die Erhebung von Antwortsicherheit können alle Wissensstufen eines Testteilnehmers gemessen werden. Dabei entspricht der Zustand von Nichtwissen der Gleichverteilung von 100%/k Antwortsicherheit auf alle Antwortoptionen, wobei k für die Anzahl der Antwortoptionen eines Items steht. Die Zuordnung von 100% Antwortsicherheit zu der richtigen Antwortoption entspricht hingegen Vollwissen (Shuford & Brown, 1975).

Die Messung von Antwortsicherheit wird in der englischsprachigen Literatur unter verschiedenen Namen beschrieben, wie beispielsweise *degrees of confidence*, *confidence marking* (Leclerq, 1983), *confidence weighting, confidence testing, probabilistic weighting, probabilistic testing* (Wood, 1991) oder *probability measurement* (Romberg, Shepler & Wilson, 1970). Im Wesentlichen können zwei Vorgehensweisen unterschieden werden: Die erste ist die Messung von sogenannten Konfidenz-

urteilen. Hierbei wählt ein Testteilnehmer bei einem Item zunächst die für ihn am wahrscheinlichsten richtige Option aus und ordnet dieser dann seine Sicherheit zu, mit der er sie tatsächlich für richtig hält. Dieses Konfidenzurteil kann zwischen 100%/k und 100% liegen. Bei z.B. vier Antwortoptionen kann es also zwischen 25% und 100% variieren. Das Konfidenzurteil kann nicht unter 25% liegen, da der Teilnehmer sich sonst für eine andere Antwortoption entschieden hätte. Die zweite Vorgehensweise ist die Messung von Antwortsicherheit. Ein Testteilnehmer verteilt dabei insgesamt 100% Antwortsicherheit auf alle angebotenen Optionen. Die Summe aller Antwortsicherheiten muss dabei je Item immer 100% oder, in Wahrscheinlichkeiten ausgedrückt, eins ergeben (De Finetti, 1965; Echternacht, 1972). Bei der Erhebung von Antwortsicherheit gibt ein Teilnehmer also zu jeder Antwortoption seine Antwortsicherheit an, während er bei der Erhebung von Konfidenzurteilen nur seine höchste Antwortsicherheit wiedergibt. Leclerq (1983) sah keine Notwendigkeit darin, die Erhebung von Konfidenzurteilen und von Antwortsicherheiten als getrennte Verfahren zu betrachten. Ben-Simon, Budescu und Nevo (1997) beurteilten die Erhebung von Konfidenzurteilen als eine vereinfachte Form der Erhebung von Antwortsicherheit. Dirkzwager (2003) postulierte die Bezeichnung Multiple Evaluation für die Messung von Antwortsicherheit. Er wählte diesen Namen, da ein Testteilnehmer nicht wie bei Multiple-Choice unter mehreren Antwortoptionen die am wahrscheinlichsten richtige auswählt, sondern alle angebotenen Antwortoptionen evaluiert. Nachfolgend wird in dieser Arbeit die Bezeichnung Multiple Evaluation für alle Verfahren zur Erhebung von Antwortsicherheit verwendet.

Eine Herausforderung bei dem Einsatz der Multiplen Evaluation besteht darin, die tatsächliche subjektive Antwortsicherheit p eines Teilnehmers zu messen, denn in einem Test kann nur die vom Teilnehmer berichtete Antwortsicherheit r beobachtet werden (Dirkzwager, 2003). Diese berichtete Antwortsicherheit muss jedoch nicht notwendigerweise mit der tatsächlichen subjektiven Antwortsicherheit des Teilnehmers übereinstimmen. Mit dem Ziel, einen Testteilnehmer zu inzentivieren, seine tatsächliche subjektive Antwortsicherheit in einem Wissenstest zu reproduzieren, damit also $p = r$ wird, wurde der Einsatz spezieller Auswertefunktionen vorgeschlagen (Murphy & Winkler, 1970). Mit diesen Funktionen beschäftigt sich der nächste Abschnitt.

4.2 Auswertefunktionen

Bei einem konventionellen Test mit Multiple-Choice erfolgt die Punktauszahlung in der Regel dichotom. Ein Teilnehmer erhält einen Punkt für eine richtige Antwort und null Punkte für eine falsche. Wie bereits ausgeführt, ist es bei einer solchen Auswertung für einen Teilnehmer die effektivste Strategie, bei Nicht- oder Teilwissen zu raten. Denn der Erwartungswert der Punkte, die er durch blindes Raten erreichen kann, beträgt pro Item 1/k Punkte, wobei k für die Anzahl der Antwortoptionen steht. Bei einem Test mit Multipler Evaluation ist die einfachste denkbare Auswertefunktion eine lineare Funktion. Bei einer linearen Auswertung wird die prozentuale Antwortsicherheit in die richtige Antwortoption als Punktwert ausgezahlt. Ordnet ein Testteilnehmer der richtigen Antwortoption also z.B. 80% Antwortsicherheit zu, so erhält er dafür 80 Punkte. Gibt er 0% Antwortsicherheit in die richtige Antwortoption an, so erhält er null Punkte. Shuford und Brown (1975) hielten eine lineare Auswertefunktion für ungeeignet, da unter einer solchen Auswertung die Testsituation zu Multiple-Choice reduziert werde. Tatsächlich ist es für den Teilnehmer eines Tests mit Multipler Evaluation bei der Auswertung durch eine lineare Funktion die beste Strategie, der plausibelsten Antwortoption immer 100% Antwortsicherheit zuzuordnen, wenn er sein Testergebnis maximieren will. Dies verdeutlicht das folgende Beispiel von zwei sich unterschiedlich verhaltenden Teilnehmern. Der eine Teilnehmer gibt zu jeder Antwortoption seine tatsächliche subjektive Antwortsicherheit ehrlich wieder. Dieser Teilnehmer wird im Folgenden als der *ehrliche* Teilnehmer bezeichnet. Der andere gibt dagegen immer 100% Antwortsicherheit in die für ihn am plausibelsten erscheinende Antwortoption an. Dieser Teilnehmer wird der *ratende* genannt. Beide Teilnehmer bearbeiten einen Test, der aus zehn Items mit je zwei Antwortoptionen besteht. Als tatsächliche subjektive Antwortsicherheit p beider Testteilnehmer gegenüber der jeweils richtigen Antwortoption wird bei jedem Item 60% angenommen. Der *ehrliche* Teilnehmer ordnet also in 60% der Fälle der richtigen Antwortoption 60% Antwortsicherheit zu. In den übrigen 40% der Fälle ordnet er dem Distraktor 60% zu und damit der richtigen Antwortoption nur 40%. Bei einer linearen Auswertung erhält er so bei zehn Items insgesamt 520 Punkte (6*60+4*40 Punkte). Der *ratende* Testteilnehmer ordnet der von ihm favorisierten Antwortoption jeweils 100% Antwortsicherheit zu. Er wählt also sechsmal die richtige Antwort und viermal eine falsche mit jeweils

100% Antwortsicherheit. Dadurch erzielt er 600 Punkte (6*100+4*0 Punkte) und damit 80 Punkte mehr als der *ehrliche* Teilnehmer. Bei einer Auswertung durch eine lineare Funktion ist es also für einen Testteilnehmer die beste Strategie, jeweils 100% Antwortsicherheit in die seiner Meinung nach am wahrscheinlichsten richtige Antwortoption anzugeben. Um einen Teilnehmer zu inzentivieren, seine tatsächliche subjektive Antwortsicherheit unverfälscht zu berichten, argumentierten Shuford, Albert und Massengill (1966, S. 1) deshalb für die Verwendung einer admissiblen Auswertefunktion:

> *Admissible probability measurement procedures utilize scoring systems with a very special property that guarantees that any student, at whatever level of knowledge or skill, can maximize his expected score if and only if he honestly reflects his degree-of-belief probabilities.*

Eine Funktion gilt demnach als admissibel, wenn sie gewährleistet, dass ein Testteilnehmer nur dann seine Punktsumme maximieren kann, wenn er seine Antwortsicherheit, also auch seine etwaige Unsicherheit, vollkommen unverfälscht berichtet. Ein Test mit Multipler Evaluation wird deshalb als reproduzierend bezeichnet, wenn die Auszahlung der Punkte durch eine admissible Auswertefunktion erfolgt (Brown & Shuford, 1973; Shuford, Albert & Massengill, 1966). Nach Toda (1963) erfüllen die Mitglieder dreier Funktionsklassen die Bedingungen admissibler Auswertefunktionen: die der sphärischen, der logarithmischen und der quadratischen Funktionsklasse. Shuford, Albert und Massengill (1966) führten einen mathematischen Beweis, in den sie bereits existierende Erkenntnisse anderer Forscher, z.B. die von Toda (1963), integrierten. Sie zeigten, dass nur logarithmische Auswertefunktionen ausschließlich auf der Basis der Antwortsicherheit, die der richtigen Antwortoption zugeordnet wurde, Punkte auszahlen. Die den Distraktoren zugeordneten Antwortsicherheiten müssen bei der Auswertung, im Gegensatz zu einer Auswertung mit einer quadratischen oder sphärischen Funktion, nicht berücksichtigt werden. Damit haben logarithmische Funktionen den Vorteil, dass die Punktauszahlung nur von der Antwortsicherheit eines Teilnehmers in die richtige Antwortoption abhängig ist, aber unabhängig von der Verteilung seiner Antwortsicherheit auf die Distraktoren (Bickel, 2007; Winkler, 1969). Logarithmische Funktionen zahlen für eine Antwortsicherheit in die richtige Antwortoption, die geringer ist als 1/k, wobei k für die Anzahl der Antwortoptionen

steht, Strafpunkte aus. Charakteristisch ist dabei, aufgrund der Nichtlinearität der Funktionen, eine höhere Strafzahlung gegenüber einer Punktauszahlung bei einer Antwortsicherheit in die richtige Antwortoption größer als 1/k (Shuford & Brown, 1975). Da eine logarithmische Funktion jedoch nur für Zahlen größer als Null definiert ist, würde die Punktauszahlung bei einer Antwortsicherheit von 0% in die richtige Antwortoption gegen -∞ Punkte streben. Um die Funktion im Testalltag einsetzen zu können, begrenzten Shuford und Brown (1975) deshalb den Definitionsbereich der Funktion auf Werte größer als null, also z.B. 1%. Den dadurch bedingten Verlust der reproduzierenden Eigenschaft für Antwortsicherheiten kleiner als 1% sahen sie als vernachlässigbar an. Ein Beispiel für eine logarithmische Auswertefunktion nach Shuford und Brown (1975) ist die folgende Funktion:

$$S(r_c) = 0{,}5*P*\log(k*r_c)$$

Dabei bedeutet:
S= Punktauszahlung
r_c= Antwortsicherheit (in Wahrscheinlichkeiten zwischen null und eins) in die richtige Antwortoption, für $r_c >= 0{,}01$
k= Anzahl der Antwortoptionen des Items
P= Auszahlungsbereich

Ist z.B. k=2 und der Auszahlungsbereich P beträgt 100 Punkte, so ergibt sich die folgende Auswertefunktion:

$$S(r_c) = 50*\log(2*r_c)$$

Abbildung 4-1 zeigt den Verlauf dieser Funktion. Im Falle völligen Nichtwissens eines Teilnehmers (50% Antwortsicherheit in beide Antwortoptionen) zahlt die Funktion null Punkte aus. Dirkzwager (2003) löste das Problem der unendlichen negativen Auszahlungen bei 0% Antwortsicherheit in die richtige Antwortoption, indem er eine modifizierte logarithmische Funktion vorschlug.

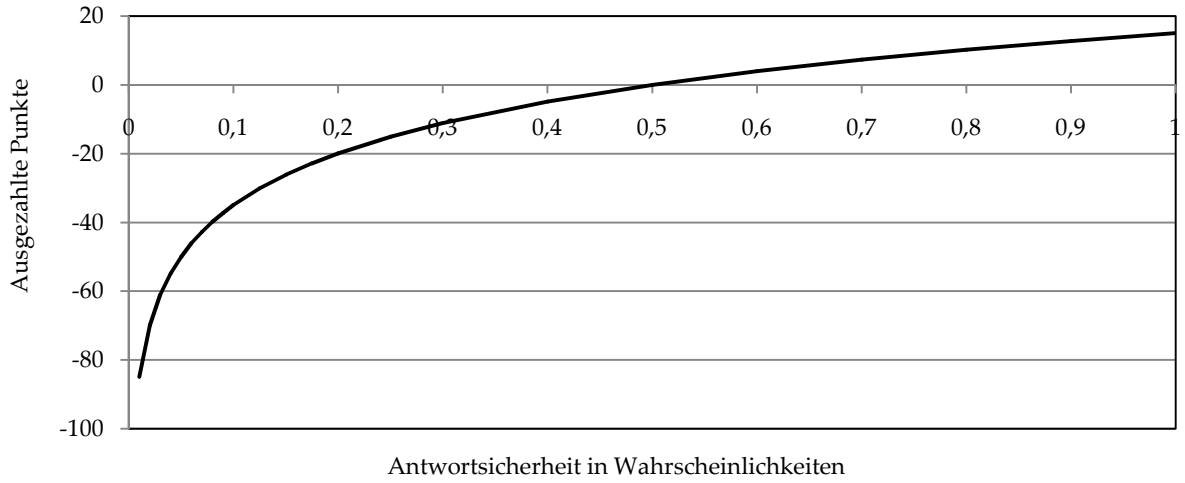

Abbildung 4-1: Logarithmische Auswertefunktion nach Shuford und Brown (1975).

Bei einer Auswertung mithilfe einer logarithmischen Funktion hat ein Testleiter die Möglichkeit, die Höhe der maximalen negativen Punktauszahlung zu bestimmen. Durch die Variation eines Toleranzfaktors T wird festgelegt, wie viele zu 100% richtig beantwortete Items erforderlich sind, um den Punkteverlust durch ein mit 0% Antwortsicherheit völlig falsch beantwortetes Item aufzuwiegen (Holmes, 2002). Ist der Toleranzfaktor beispielsweise T=3, so werden einem Teilnehmer bei 0% Antwortsicherheit in die richtige Antwortoption 300 Punkte abgezogen. Für ein zu 100% richtig beantwortetes Item erhält er 100 Punkte gutgeschrieben. Er benötigt also drei vollkommen richtig beantwortete Items, um von einem Punktestand von -300 Punkten auf 0 Punkte zu gelangen. Dieser Punktestand würde als Gesamtergebnis jedoch völliges Nichtwissen bedeuten. Damit geht ein Teilnehmer also, wenn er seine Antwortsicherheit nicht seiner tatsächlichen subjektiven Antwortsicherheit entsprechend berichtet, ein hohes Risiko ein, seinen Punktestand erheblich zu reduzieren. Der Toleranzfaktor T kann variiert werden, je nachdem wie konsequent nur geringe Antwortsicherheiten in die richtige Antwortoption sanktioniert werden sollen. Abbildung 4-2 zeigt den Verlauf der Auswertefunktion nach Dirkzwager (2003) für vier verschiedene Toleranzfaktoren und Items mit zwei Antwortoptionen. Um den gewünschten Toleranzfaktor T in der Auswertung zu realisieren, wird in der von Dirkzwager (2003) vorgeschlagenen Auswertefunktion der Toleranzparameter t variiert. Je kleiner t gewählt wird, desto

höher sind die möglichen Strafzahlungen. Hohe Strafzahlungen sollen die Testteilnehmer inzentivieren, ihre tatsächliche subjektive Antwortsicherheit

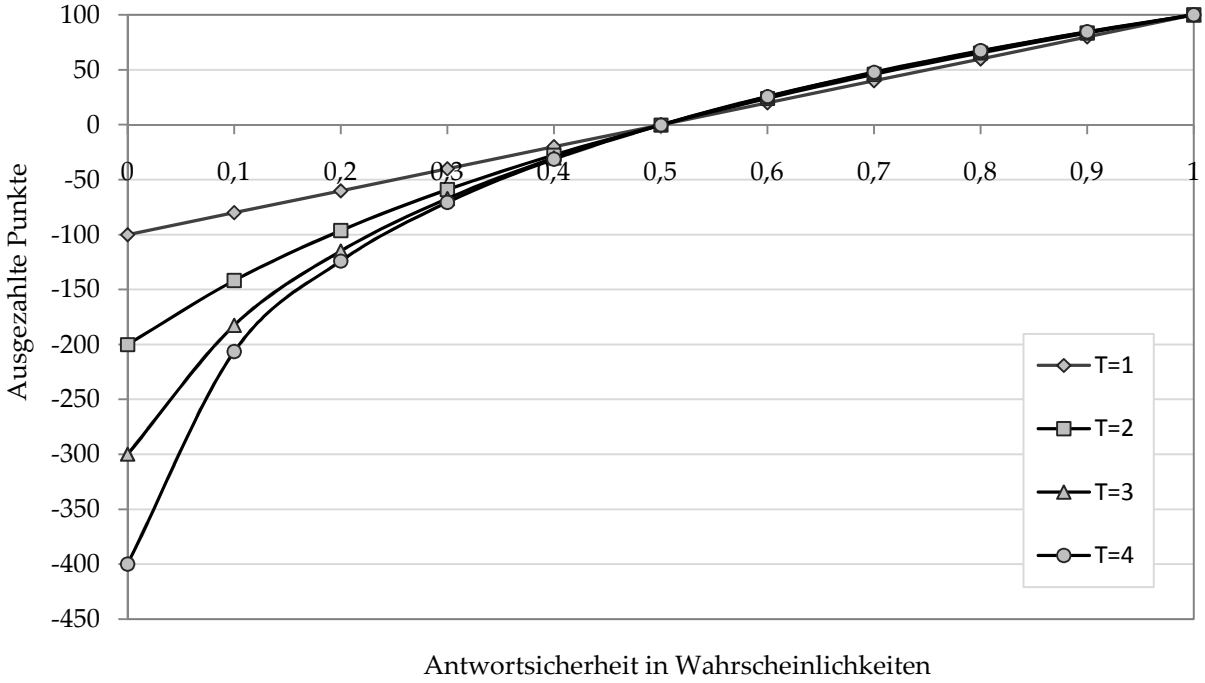

Abbildung 4-2: Logarithmische Auswertefunktion nach Dirkzwager (2003) für verschiedene Toleranzfaktoren T für Items mit zwei Antwortoptionen.

unverfälscht zu berichten, und zwar umso mehr, je höher diese Strafzahlungen sind. Nähert t sich seinem Maximum 1/k an, wobei k für die Anzahl der Antwortoptionen steht, dann wird eine nur geringe Antwortsicherheit in die richtige Antwortoption nicht mehr durch hohen Punktabzug bestraft, und Multiple Evaluation nähert sich Multiple-Choice mit Ratekorrektur an (Dirkzwager, 2003, S. 338 f.):

> *When the tolerance parameter t approaches its maximum 1/k, the multiple evaluation method approaches multiple choice with the so-called correction for guessing. The expected score is equal to or larger than zero for an unknowing participant and examinees maximize their possible gain by guessing. The smaller the tolerance parameter t, the more guessing is discouraged and realistic probability estimates are furthered; thus, accurate knowledge measurement with the possibility of discovering fallacies is better possible.*

Die logarithmische Auswertefunktion nach Dirkzwager (2003) lautet:

$$S\left(r_c(i)\right) = \frac{\ln[(1-t*k)*r_c(i)+t] + \ln(k)}{\ln(1-t*k+t) + \ln(k)}$$

Dabei bedeutet:

S= Auszahlung

r_c=Antwortsicherheit, der richtigen Alternative zugeordnet

t= Toleranzparameter

k= Anzahl der Antwortoptionen

i= Item

Der Toleranzfaktor T steht in der folgenden Beziehung zu t und k (Holmes, 2002):

$$T(t,k) = - \frac{\ln(t) + \ln(k)}{\ln(1-t*k+t) + \ln(k)}$$

Dabei bedeutet:

T=Toleranzfaktor

t= Toleranzparameter

k= Anzahl der Antwortoptionen

Tabelle 4-1

Toleranzparameter t(T,k) für verschiedene Anzahlen von Antwortoptionen (k) und Toleranzfaktoren (T), (Holmes, 2002, S. 44).

	k=2	k=3	k=4	k=5
T=1	0,4999	0,1667	0,0833	0,0500
T=2	0,1910	0,0447	0,0174	0,0086
T=3	0,0804	0,0134	0,0041	0,0016
T=4	0,0362	0,0043	0,0010	0,0003

Tabelle 4-1 zeigt für vier Toleranzfaktoren T die entsprechenden Werte des Toleranzparameters t.

Für die logarithmische Auswertung nach Dirkzwager (2003) wird wiederum das Beispiel zweier sich unterschiedlich verhaltender Teilnehmer betrachtet. Der *ehrliche* Teilnehmer berichtet seine Antwortsicherheit vollkommen unverfälscht, während der *ratende* immer 100% Antwortsicherheit in die ihm am plausibelsten erscheinende Antwortoption angibt. Für beide Teilnehmer wird wieder eine tatsächliche subjektive Antwortsicherheit von 60% angenommen. Betrachtet werden die Punktsummen von zehn Items mit je zwei Antwortoptionen. Tabelle 4-2 zeigt diese Punktsummen für die Auswertung mit den Toleranzfaktoren T=1, T=2, T=3 und T=4. Die Berechnung der Punktsummen wird im Folgenden anhand der

Tabelle 4-2

Summe der Auszahlungen für verschiedene Toleranzfaktoren für zehn Items mit zwei Antwortoptionen für den *ehrlichen* und den *ratenden* Testteilnehmer.

Toleranzfaktor	T=1	T=2	T=3	T=4
ehrlicher Teilnehmer	40	36	30	32
ratender Teilnehmer	200	-200	-600	-1000

logarithmischen Auswertefunktion mit einem Toleranzfaktor von T=4 demonstriert. Der *ratende* Testteilnehmer ordnet immer der von ihm favorisierten Antwortoption jeweils 100% Antwortsicherheit zu. Er wählt also sechsmal die richtige Antwort mit jeweils 100% Antwortsicherheit. Dadurch erzielt er 600 Punkte (6*100 Punkte). Jedoch wählt er auch viermal einen Distraktor mit einer Antwortsicherheit von 100% und dafür werden ihm 1600 Punkte [4*(-400 Punkte)] abgezogen. Sein Gesamtpunktestand beträgt deshalb -1000 Punkte. Der *ehrliche* Teilnehmer ordnet in 60% der Fälle der richtigen Antwortoption 60% Antwortsicherheit zu und erhält dafür 156 Punkte (6*26 Punkte). In den übrigen 40% der Fälle ordnet er dem Distraktor 60% zu und damit der richtigen Antwortoption nur 40%. Dafür werden ihm -124 Punkte [4*(-31 Punkte)] abgezogen. Insgesamt erhält der *ehrliche* Teilnehmer 32 Punkte und damit 968 mehr als der *ratende*. Der Vergleich zeigt für Items mit zwei Antwortoptionen, dass wenn der Toleranzfaktor T=1 beträgt, es die beste Strategie für einen Teilnehmer ist, immer 100% Antwortsicherheit in die plausibelste

Antwortoption anzugeben. Bei einem Toleranzfaktor T>1 hingegen ist die unverfälschte Wiedergabe der tatsächlichen subjektiven Antwortsicherheit eines Teilnehmers die einzige Strategie, wenn er seine Punktsumme maximieren will. Mithilfe von Strafzahlungen soll ein Teilnehmer also inzentiviert werden, sein Wissen unverfälscht zu reproduzieren. Mithilfe einer logarithmischen Auswertung soll so der wahre Wissensstand eines Teilnehmers gemessen und die Varianz der beobachtbaren Testwerte im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren verringert werden können (Dirkzwager, 2003). Theoretisch stellt die Auswertung mit einer admissiblen logarithmischen Funktion damit eine geeignete Lösung dar, um die Güte eines Tests zu verbessern (Hogarth, 1975). Offen ist jedoch, ob eine solche Verbesserung auch empirisch gezeigt werden kann (Keren, 1991; Toda, 1963). Die Auszahlung der Punkte anhand einer logarithmischen Funktion könnte, besonders für Testteilnehmer ohne gutes mathematischen Verständnis, nicht einfach zu verstehen sein, was ihren Einsatz im Testalltag erschweren würde (Ben-Simon, Budescu & Nevo, 1997; Hogarth, 1975). Hogarth (1975) führte das wichtige Argument an, dass der Theorie über die Auswertung mithilfe admissibler Funktionen die Annahme zugrunde liege, dass Testteilnehmer ihre tatsächliche subjektive Antwortsicherheit kennen würden, was in der Regel aber nicht der Fall sei.

In den Experimenten dieser Arbeit wurden Funktionen der von Dirkzwager (2003) modifizierten logarithmischen Auswertefunktionsklasse zur Auszahlung der Punkte eingesetzt. Es wurde geprüft, ob ein Test mit Multipler Evaluation mit einer logarithmischen Auswertung eine höhere Reliabilität und eine höhere Validität zeigt als mit einer linearen. Als eine weitere Fragestellung wurde der Einfluss der Höhe des Toleranzfaktors in der logarithmischen Auswertung, und damit der möglichen Strafzahlungen, auf die Reliabilität und die Validität eines Tests untersucht.

4.3 Güte der Kalibrierung von Testteilnehmern

Koriat, Lichtenstein und Fischhoff (1980) sahen es als eine bemerkenswerte Fähigkeit des menschlichen Gedächtnisses an, über Wissen über den eigenen Wissensstand zu verfügen. Ein gut bekanntes Phänomen ist es jedoch auch, dass Menschen bei vielen Denkprozessen zur *Overconfidence* neigen (Metcalf, 1998). Kalibrierungsexperimente,

die mithilfe von Konfidenzurteilen durchgeführt wurden, betätigten, dass in der Regel nicht alle Teilnehmer eines Tests perfekt kalibriert sind, sondern dass ein Teil der Teilnehmer *Overconfidence* zeigt (Fischhoff, Slovic & Lichtenstein, 1977; Keren, 1991; Koriat, Lichtenstein & Fischhoff, 1980). Offen ist, warum manche Testteilnehmer von einer perfekten Kalibrierung abweichen, während andere gut kalibriert sind (Shuford & Brown, 1975). Eine mögliche Ursache für eine mangelnde Kalibrierung könnte sein, dass Teilnehmer ihr Gefühl der Unsicherheit nicht in akkurate numerische Antwortsicherheiten übersetzen können und daher unangemessene Antwortsicherheiten berichten (Adams & Adams, 1961). Das Problem wird noch verstärkt, wenn Testteilnehmer daran gewöhnt sind, nur dichotome Antworten auf Multiple-Choice-Items zu geben und daher keine Übung darin haben, ihre subjektive Antwortsicherheit differenziert anzugeben (Holmes, 2002). Die Aufgabenschwierigkeit kann zudem die Güte der Kalibrierung eines Teilnehmers beeinflussen (Keren, 1991; Lichtenstein & Fischhoff, 1980). Bei schwierigen Aufgaben zeigen Teilnehmer häufig *Overconfidence*, bei einfachen Aufgaben eher *Underconfidence*. Dieser Effekt wird als *Hard-Easy-Effekt* bezeichnet (May, 1987). Lichtenstein und Fischhoff (1980) fanden Hinweise darauf, dass die Reihenfolge, in der einfache und schwierige Items präsentiert werden, ebenfalls einen Einfluss auf die Güte der Kalibrierung hat. Bei einem abrupten Wechsel von einem Set einfacher zu schwierigen Items ist demnach eine Verschlechterung der Kalibrierung zu erwarten. Empirische Befunde deuten auch darauf hin, dass die Güte der Kalibrierung der Teilnehmer von der Art des Testmaterials beeinflusst wird (Jans & Leclercq, 1997; Keren, 1991; May, 1987). Keren (1991) beschrieb *Overconfidence* besonders bei Items, die Allgemeinwissen prüfen.

Die empirische Befundlage gibt jedoch auch Hinweise darauf, dass die Kalibrierung eines Testteilnehmers durch Training unter *Feedback* verbessert werden kann. Lichtenstein und Fischhoff (1980) konnten anhand von Konfidenzurteilen beispielsweise zeigen, dass eine Trainingseinheit von 200 Items, gefolgt von intensivem *Feedback*, ausreicht, um die Kalibrierung von Teilnehmern zu verbessern. Ein Drittel der Teilnehmer wies dabei bereits vor dem Training eine gute Kalibrierung auf. Auch Rippey und Voytovich (1982) zeigten, dass Testteilnehmer in der Lage sind, die Güte ihrer Kalibrierung über eine Serie von Tests hinweg zu verbessern. Rippey und Voytovich (1982) führten die Tests mit Multipler Evaluation

und einer logarithmischen Auswertefunktion (Shuford, Albert & Massengill, 1966) durch. Die Teilnehmer erhielten ein *Feedback* über die Güte ihrer Kalibrierung nach jedem Test. Die Ergebnisse dieser Studien lassen jedoch nicht generell den Schluss zu, dass die Testteilnehmer wirklich ihre Kalibrierung dauerhaft verbessern konnten. So kritisierte beispielsweise May (1987), dass es zweifelhaft sei, ob aus einer gefundenen Verbesserung mehr geschlossen werden könne, als dass die Testteilnehmer ihre Antwortsicherheiten etwas herabgesetzt hätten.

Zusammenfassend ist also davon auszugehen, dass ein unbekannter Teil der Teilnehmer eines Tests nicht perfekt kalibriert ist. Offen ist, ob die Güte der Kalibrierung eines Teilnehmers durch Training dauerhaft verbessert werden kann. Zudem wäre die Durchführung eines Trainings, das wahrscheinlich für jedes Testmaterial spezifisch durchgeführt werden müsste, sehr aufwendig und nicht immer möglich. Um dieses Problem zu lösen, wurde mit dem Ziel, die Fehlkalibrierung der Teilnehmer in einem Test nachträglich auszugleichen, eine Korrektur der berichteten Antwortsicherheiten auf der Basis eines individuellen Realismusindex vorgeschlagen (Brown & Shuford, 1973; Holmes, 2002). Die Bestimmung dieses individuellen Realismusindex und die Korrektur der berichteten Antwortsicherheiten auf der Basis dieses Index werden in den nächsten beiden Abschnitten vorgestellt.

4.3.1 Individueller Realismusindex als Maß der Güte der Kalibrierung

Bei einem Test mit Multipler Evaluation wird eine lineare Beziehung zwischen der tatsächlichen subjektiven Antwortsicherheit eines Teilnehmers p und seiner berichteten Antwortsicherheit r zu Grunde gelegt. Entspricht die in die richtige Antwortoption berichtete Antwortsicherheit r_c eines Teilnehmers seiner tatsächlichen subjektiven Antwortsicherheit p_c, ist also $p_c=r_c$, dann gilt er als perfekt kalibriert. Ist $r_c \neq p_c$, dann weicht er von einer perfekten Kalibrierung ab. Bei einem Test mit Multipler Evaluation wird die Beziehung zwischen p_c und r_c und damit die Güte der Kalibrierung eines Teilnehmers anhand seines individuellen Realismusindex mithilfe der folgenden Formel bestimmt (Brown & Shuford, 1973; Holmes, 2002, S. 51):

$$p_c = a*r_c+(1-a)/k$$

Dabei bedeutet:

p_c= tatsächliche subjektive Antwortsicherheit eines Teilnehmers in die richtige Antwortoption (in Wahrscheinlichkeiten von null bis eins)

a= Realismusindex (Steigung der Geraden)

r_c= berichtete Antwortsicherheit eines Teilnehmers in die richtige Antwortoption (in Wahrscheinlichkeiten von null bis eins)

k= Anzahl der Antwortoptionen eines Items

Der Realismusindex a ist die Steigung dieser Geraden und kann mithilfe der folgenden Formel ermittelt werden (Brown & Shuford, 1973; Holmes, 2002, S. 132):

$$a = \frac{k^* \sum_{i=1}^{m} r_c(i) - m}{k^* \sum_{i=1}^{m} \sum_{j=1}^{k} r(i,j)^2 - m}$$

Dabei bedeutet:

m= Anzahl der Items

k= Anzahl der Antwortoptionen

r_c= berichtete Antwortsicherheit eines Teilnehmers in die richtige Antwortoption

r= Antwortsicherheiten in alle Antwortoptionen

i= Item, i=1, …, m

j= Antwortoptionen des Items, j=1, …, k

Nimmt a den Wert eins an, dann ist ein Testteilnehmer perfekt kalibriert, denn r_c=p_c. Ist a kleiner als eins, dann tendiert er dazu, höhere Antwortsicherheiten zu berichten, als es einer tatsächlichen subjektiven Antwortsicherheit entspricht, er zeigt *Overconfidence*. Nimmt a einen Wert größer als eins an, dann berichtet der Testteilnehmer häufig geringere Antwortsicherheiten als es seiner tatsächlichen

subjektiven Antwortsicherheit entspricht und zeigt damit *Underconfidence* (Brown & Shuford, 1973; Dirkzwager, 2003; Holmes, 2002; Shuford & Brown, 1975). Abbildung *4-3* zeigt jeweils die Gerade für einen Teilnehmer, der *Overconfidence* zeigt, mit einem Realismusindex von a=0,75, und einen Teilnehmer, der *Underconfidence* zeigt, mit einem Realismusindex von a=1,2 im Vergleich zu einem perfekt kalibrierten Testteilnehmer mit einem Realismusindex von a=1 (Holmes, 2002).

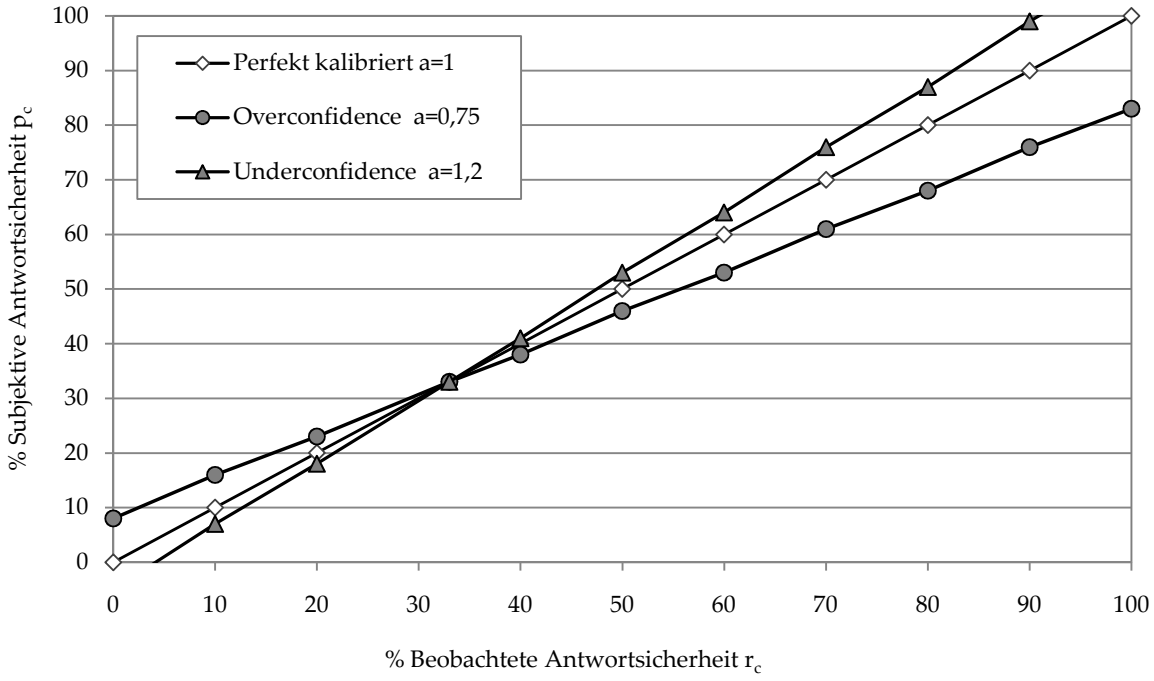

Abbildung 4-3: Kalibrierungsgeraden bei verschiedenen Werten für den Realismusindex a.

4.3.2 Korrektur von Antwortsicherheiten anhand eines individuellen Realismusindex

Bei einer Korrektur auf der Basis des individuellen Realismusindex können zwei Ziele verfolgt werden. Erstens die Kalibrierung der Testteilnehmer auf lange Sicht zu verbessern. Zweitens die Varianz der beobachtbaren Testwerte, die durch eine verfälschte Wiedergabe der tatsächlichen subjektiven Antwortsicherheit erzeugt wird, zu reduzieren. Für die erste Anwendungsmöglichkeit kann der ermittelte

Realismusindex dazu genutzt werden, den Teilnehmer über seine Abweichung von einer perfekten Kalibrierung zu informieren, um seinen Lernprozess zu unterstützen. Ein perfekt kalibrierter Teilnehmer ist in der Lage, seinen Wissensstand richtig einzuschätzen und weiß daher, was er lernen muss. Bei einem nicht perfekt kalibrierten Teilnehmer hingegen ist diese Einschätzung fehlerhaft. Eine Rückmeldung des Realismusindex könnte ihm möglicherweise helfen, diese Einschätzung zu verbessern (Jans & Leclercq, 1997). Die Rückmeldung des Realismusindex als eine abstrakte Zahl liefert einem Testteilnehmer jedoch wenig nutzbare Information. Eine bessere Vorgehensweise ist es daher, dem Teilnehmer anhand der Punkte, die er aufgrund seines Antwortverhaltens verloren hat, seine verzerrte Kalibrierung zu visualisieren. Dazu werden seine Antwortsicherheiten nachträglich auf der Basis seines individuellen Realismusindex korrigiert. Durch die Korrektur erhält ein Teilnehmer die Antwortsicherheiten, die er hätte wiedergeben müssen, wenn er im Test perfekt kalibriert gewesen wäre. Die vom Teilnehmer berichteten Antwortsicherheiten werden also seinen tatsächlichen subjektiven angenähert. Diese korrigierten Antwortsicherheiten des Testteilnehmers werden mithilfe der folgenden Formel ermittelt (Holmes, 2002, S. 132):

$$r_{korrigiert} = a*r_c+(1-a)/k$$

Dabei bedeutet:

$r_{korrigiert}$= realismusindexkorrigierte Antwortsicherheit in die richtige Antwort (in Wahrscheinlichkeiten von null bis eins)

a= individueller Realismusindex des Testteilnehmers

r_c= berichtete Antwortsicherheit in die richtige Antwort (in Wahrscheinlichkeiten von null bis eins)

k= Anzahl der Antwortoptionen

Ist ein Testteilnehmer perfekt kalibriert, d.h. sein Realismusindex beträgt a=1, so werden seine Antwortsicherheiten mit der Formel $r_{korrigiert}=1*r_c+0$ korrigiert. Diese Korrektur verändert die berichtete Antwortsicherheit r_c nicht, denn bei einem perfekt kalibrierten Teilnehmer entspricht diese seiner tatsächlichen subjektiven Antwort-

sicherheit in die richtige Antwort, und damit ist $r_{korrigiert} = r_c$. Bei einem nicht perfekt kalibrierten Testteilnehmer, der *Overconfidence* zeigt, und daher beispielsweise einen Realismusindex von a=0,75 erzielt, würde bei Items mit drei Antwortoptionen die berichtete Antwortsicherheit in die jeweils richtige Antwort mit der Formel $r_{korrigiert}=0,75*r_c+0,083$ korrigiert, bei einem Teilnehmer, der *Underconfidence* zeigt, mit z.B. a=1,2 durch die Formel $r_{korrigiert}=1,2*r_c-0,066$. Tabelle 4-3 zeigt für beide Fälle die prozentualen Antwortsicherheiten in die richtigen Antworten für zehn Items und die durch eine logarithmische Funktion nach Dirkzwager (2003) mit dem Auszahlungsbereich -300 bis 100 (Toleranzfaktor T=3) dafür ausgezahlten Punkte vor und nach der Korrektur. Der Teilnehmer mit *Overconfidence* würde ohne die Korrektur eine Punktsumme von -219 Punkten, der Teilnehmer mit *Underconfidence* von 811 Punkten erhalten. Bei einem Teilnehmer mit *Overconfidence* und einem Realismusindex von a=0,75 werden nur geringe Antwortsicherheiten in die richtige

Tabelle 4-3

Prozentuale Antwortsicherheiten in die richtige Antwort eines Teilnehmers mit *Underconfidence* und eines Teilnehmers mit *Overconfidence* bei der Auswertung mit einer logarithmischen Funktion nach Dirkzwager (2003) mit einem Auszahlungsbereich von -300 bis 100 Punkten vor und nach der Korrektur auf der Basis des individuellen Realismusindex.

Teilnehmer		Item-Nr	1	2	3	4	5	6	7	8	9	10	Summe
Over-confident a=0,75	vor der Korrektur	% Antwort-sicherheit	0	0	30	33	50	50	50	80	100	100	
		ausgezahlte Punkte	-300	-300	-9	0	37	37	37	79	100	100	-219
	nach der Korrektur	% Antwort-sicherheit	8	8	31	33	46	46	46	68	83	83	
		ausgezahlte Punke	-119	-119	-7	0	29	29	29	65	83	83	73
Under-confident a=1,2	vor der Korrektur	% Antwort-sicherheit	60	60	70	80	80	90	90	100	100	100	
		ausgezahlte Punkte	53	53	67	79	79	90	90	100	100	100	811
	nach der Korrektur	% Antwort-sicherheit nicht begrenzt	65	65	77	89	89	101	101	113	113	113	
		% Antwort-sicherheit begrenzt auf 0 bis 100%	65	65	77	89	89	100	100	100	100	100	
		ausgezahlte Punke	60	60	76	89	89	100	100	100	100	100	874

Anzahl der Antwortoptionen=3.

Antwortoption von z.B. 0% auf 8% nach oben korrigiert. Daher werden z.B. statt 300 Punkten nur 119 Punkte je Item abgezogen. Alle Antwortsicherheiten werden durch die Korrektur einem Wert von 100%/k,wobei k für die Anzahl der Antwortoptionen steht, bei drei Antwortoptionen also gerundet 33%, angenähert. Daher reduziert sich im Gegenzug auch die Auszahlung von 100 Punkten, die der Teilnehmer für eine zu 100% richtige Antwort erhalten hatte, auf 83 Punkte. Durch die Korrektur erhält der Testteilnehmer mit *Overconfidence* insgesamt 73 Punkte, und damit 292 Punkte mehr, ausgezahlt. Bei einem Teilnehmer mit *Underconfidence* mit einem Realismusindex von a>1 kann die Korrektur zu korrigierten Antwortsicherheiten kleiner als 0% oder größer als 100% führen. In diesem Fall müssen die prozentualen Antwortsicherheiten auf den Wertebereich von 0% bis 100% begrenzt werden (Holmes, 2002). Die korrigierte Punktsumme für den Teilnehmer mit *Overconfidence* mit einem Realismusindex von a=1,2 beträgt demnach 874 Punkte. Er erhält durch die Korrektur also 63 Punkte mehr. Die Rückmeldung des Vergleichs zwischen den tatsächlich ausgezahlten Punkten und den Punkten, die der Teilnehmer bei einer perfekt kalibrierten Wiedergabe seiner tatsächlichen subjektiven Antwortsicherheit hätte erreichen können, sollen ihm helfen, seine Kalibrierung zu verbessern. Da die Berechnung des Realismusindex nur über möglichst viele Items hinweg sinnvoll ist, kann die Rückmeldung der Punktedifferenz jedoch erst nach Abschluss eines Tests erfolgen.

Wie ausgeführt, werden bei der Korrektur die Antwortsicherheiten eines Teilnehmers auf der Basis seines individuellen Realismusindex korrigiert. Die erneute Punktauszahlung erfolgt dann auf der Basis der Antwortsicherheiten, die ein Teilnehmer hätte angeben müssen, wenn er perfekt kalibriert gewesen wäre. Die Varianz der beobachtbaren Testwerte, die darin begründet ist, dass ein Teilnehmer seine subjektive Antwortsicherheit verfälscht berichtet, kann so nachträglich reduziert werden. Es war daher anzunehmen, dass mithilfe dieses Korrekturverfahrens die Güte eines Tests mit Multipler Evaluation verbessert werden kann. Rippey und Voytovich (1983) zeigten empirisch eine Verbesserung der Reliabilität und der Validität eines Tests aufgrund dieser Korrektur. In allen Experimenten dieser Arbeit wurde eine Korrektur der erhobenen Antwortsicherheit auf der Basis des individuellen Realismusindex durchgeführt. Es wurde geprüft, ob

diese Korrektur ein geeignetes Verfahren ist, um die Reliabilität und die Validität eines Tests mit Multipler Evaluation nachträglich zu verbessern.

4.4 Information über die Punktauszahlung

Dirkzwager (1993) sowie Brown und Shuford (1973) nahmen an, dass bei Multipler Evaluation die Information des Teilnehmers über die Punktauszahlung, während dieser ein Item bearbeitet, eine entscheidende Rolle für das Antwortverhalten der Teilnehmer spielt (Brown & Shuford, 1973, S. 8):

> *In our experience it is absolutely essential that the person assigning probabilities focus upon the conditional scores as provided by a reproducing scoring system, rather than on the probabilities themselves. If for some reason the person ignores the scores and focuses upon the probabilities, the resulting probability assignments will almost certainly be biased and there will be a loss of potential information.*

Ein *Feedforward* soll den Testteilnehmer in dem Moment, in dem er seine Antwortsicherheit berichtet, darüber informieren, welche Konsequenzen seine Antwort in Form von ausgezahlten Punkten haben kann (Anderson, 1982; Brown & Shuford, 1973; Dirkzwager, 2003). Trifft ein Teilnehmer seine Entscheidung nicht auf der Basis der zu erwartenden Punktauszahlung, sondern auf der Basis der prozentualen Antwortsicherheiten, könnte dies nach Brown und Shuford (1973) zu einer Verzerrung der Antwortsicherheiten führen. Der Teilnehmer nehme dann seine Antwortsicherheiten als eine Messung seiner Leistung wahr. Die optimale Strategie, um seine Leistung zu maximieren, wäre es dann, der am wahrscheinlichsten richtigen Antwortoption eine hohe Antwortsicherheit zuzuweisen. Unmittelbar nach der Bearbeitung eines Items soll der Teilnehmer *Feedback* darüber erhalten, welche Antwortoption tatsächlich richtig war und wie viele Punkte er verloren oder gewonnen hat. Bei Multipler Evaluation mit einer logarithmischen Auswertung soll die Information über die Punktauszahlung, die aus dem *Feedforward* der zu erwartenden Punktauszahlung und dem *Feedback* der tatsächlich ausgezahlten Punkte besteht, den Teilnehmer inzentivieren, seine tatsächlichen subjektiven Antwortsicherheiten vollkommen unverfälscht zu berichten (Dirkzwager, 2003). Als Folge daraus soll die Varianz der beobachtbaren Testwerte eines Tests, die durch

eine verfälschte Reproduktion der tatsächlichen subjektiven Antwortsicherheiten hervorgerufen wird, reduziert werden können.

Sharp, Cutler und Penrod (1988) untersuchten die Wirkung von *Feedback* auf die Kalibrierung von Testteilnehmern. Sie untersuchten 54 Teilnehmer in vier Tests durch eine Erhebung von Konfidenzurteilen. Nach jedem Test erhielten die Testteilnehmer als *Feedback* ihre Konfidenzurteile und die Häufigkeit, mit der diese Urteile richtig waren. Die Hypothese, dass *Feedback Overconfidence* reduziere und die Kalibrierung der Testteilnehmer verbessere, konnten sie nicht belegen. Hierbei ist jedoch kritisch zu sehen, dass die Teilnehmer keine direkte Rückmeldung nach jedem Item erhielten, sondern erst am Ende des Tests. Außerdem verwendeten Sharp, Cutler und Penrod (1988) keine admissible Auswertefunktion. Delgado und Prieto (2003) untersuchten den Einfluss von *Feedback* mit Multiple-Choice-Tests an 240 Testteilnehmern. Sie verglichen zwei Auswertebedingungen. In der einen erhielten die Teilnehmer als Auszahlung die Anzahl der richtigen Antworten, in der anderen wurden falsche Antworten durch Punktabzug bestraft. In beiden Auswertebedingungen erhielt die eine Gruppe von Teilnehmern *Feedback* und die andere nicht. Das *Feedback* bestand aus einem hohen Ton, wenn die Antwort richtig war, und einem tiefen, wenn sie falsch war. Delgado und Prieto (2003) fanden keinen Einfluss von *Feedback* auf die Testwerte. In den Bedingungen mit *Feedback* beobachteten sie eine Verschlechterung der Reliabilität. Geschlechtsunterschiede zeigten sich dabei nicht. Kulhavy, Yekovich und Dyer (1976) beobachteten, dass der Nutzen, den ein Teilnehmer aus einem *Feedback* zieht, davon abhängig sein kann, ob das Item richtig gelöst wurde. Sie zeigten durch Tests mit Konfidenzurteilen, dass Testteilnehmer *Feedback* am intensivsten studieren, wenn ihre Antwortsicherheit in eine falsche Option hoch ist. Die Teilnehmer, die *Feedback* erhielten, lösten zudem in einem Nachtest signifikant mehr Items richtig als die Teilnehmer der Bedingung ohne *Feedback*. Empirische Ergebnisse deuten außerdem darauf hin, dass *Feedback* die Leistung eines Teilnehmers auch negativ beeinflussen kann. Beckmann und Beckmann (2005) vermuteten, dass der Grad der Übereinstimmung des *Feedbacks* mit dem akademischen Selbstkonzept einer Person eine wichtige Rolle dabei spiele, ob *Feedback* eine positive oder negative Wirkung auf die Testperformanz hat. In einem Experiment konnten sie zeigen, dass die Gabe von *Falsch-Richtig-Feedback* nicht hilfreich ist, wenn das *Feedback* keine weiteren Informationen enthält. Es kann

Besorgnis auslösen und so die aufgabenrelevante Informationsverarbeitung stören. Dirkzwager (2003, S. 351) fand Hinweise darauf, dass ein direktes *Ergebnis-Feedback* Teilnehmer auch zum Raten animieren kann, anstatt sie vom Raten abzuhalten. Er verwendete eine Computeranwendung, um einen Test mit reproduzierender Multipler Evaluation mit elfjährigen Schülern durchzuführen und beobachtete, dass viele den Test wie ein Spiel betrachteten:

> *We have, however, some indication that this immediate feedback turns the computer into a slot machine and might encourage gambling because participants think risking large losses and claiming large gains when they see immediately whether they have gained or lost the bet is exciting. This seems to overpower the actual loss that, when it is very large, is taken as an event that enhances the fun of the game.*

Die Visualisierung der zu erwartenden Punktauszahlung in dem Moment, in dem ein Teilnehmer seine Antwortsicherheit berichtet, sowie ein *Feedback* über die tatsächlich ausgezahlten Punkte soll einen Teilnehmer inzentivieren, seine tatsächliche subjektive Antwortsicherheit gemäß seinem Wissen unverfälscht zu berichten. Vor dem Hintergrund der hier dargelegten Befunde lässt sich die Wirkung einer Auszahlungsinformation aber nicht eindeutig ableiten. Es ist anzunehmen, dass *Feedback* die Aufmerksamkeit des Teilnehmers bei hohen Antwortsicherheiten in eine falsche Antwortoption erhöht, besonders wenn ein Abzug von Punkten erfolgt ist. Die Rückmeldung von Strafzahlungen könnte aber auch negative emotionale Reaktionen wie Besorgnis oder Frustration auslösen (Bodemann, Perrez, Schär & Trepp, 2004). Dies könnte eine Erhöhung der Varianz der beobachtbaren Testwerte zur Folge haben und dann wäre es auch denkbar, dass eine Information über die Auszahlung zu einer Verschlechterung der psychometrischen Qualität eines Tests führt. Die unmittelbare Rückmeldung des Gewinns oder Verlusts von Punkten kann einen Testteilnehmer zudem dazu animieren, die Punkte als eine Art Wetteinsatz zu betrachten. Als eine Folge daraus könnte eine verfälschte Reproduktion der tatsächlichen subjektiven Antwortsicherheiten eher inzentiviert statt reduziert werden. Ein Ziel der vorliegenden Arbeit war es deshalb, zu klären, ob durch eine Information des Teilnehmers über die Auszahlung die Güte eines Tests mit Multipler Evaluation verbessert werden kann.

5 Empirische Untersuchungen verschiedener Antwortverfahren und Auswertefunktionen

Obwohl schon in der Mitte des 20. Jahrhunderts die theoretischen Grundlagen entwickelt wurden, stehen bisher zu wenige empirische Daten zur Verfügung, um eine zuverlässige Einschätzung der psychometrischen Eigenschaften eines Tests mit Multipler Evaluation zu erlauben (Hambleton, Roberts & Traub, 1970; Holmes, 2002). Dirkzwager (1993, S. 165) beklagte, dass Forscher es aufgegeben haben, das Antwortverfahren Multiple Evaluation zu untersuchen:

> *The literature on proper scoring rules seems to have died out. Around the year 1980 we find quite a few scientific publications on the topic and the strong arguments to apply them to educational testing are well known, be it mainly among the scientifically trained supporters, but not among the educators and policy makers in education. They are quite happy with Multiple-Choice-Tests, they rather forget the arguments raised against them when they were first introduced on a large scale, and they resist discussing a paradigm that is new to them. Supporters and researchers of the methods based on proper scoring rules seem to be frustrated and leave the topic. That is regrettable. Many interesting questions remain for study, research and development …*

Bei der folgenden Betrachtung empirischer Befunde wird besondere Aufmerksamkeit auf die jeweils verwendete Auswertefunktion und das Verfahren, durch das Antwortsicherheit erhoben wurde, gerichtet. Die Gliederung des Abschnitts erfolgt nach dem verwendeten Antwortverfahren.

5.1 Rangreihenverfahren

Bei Rangreihenverfahren ordnen Testteilnehmer alle Antwortoptionen eines Items in Abhängigkeit davon, mit welcher Wahrscheinlichkeit sie diese als richtig einschätzen (De Finetti, 1965). Diamond (1975) untersuchte ein Rangreihenverfahren im Vergleich zu einem herkömmlichen Multiple-Choice-Test in einer Serie von Experimenten. Das Testmaterial bestand aus Items mit vier Antwortoptionen. An dieser Studie nahmen 84 Studenten teil. Die Teilnehmer erhielten als Punktauszahlung drei Punkte, wenn sie der richtigen Antwort Rang eins zugewiesen

hatten, zwei Punkte bei Rang zwei, einen Punkt bei Rang drei und null Punkte bei Rang vier. Diamond (1975) fand eine leicht verbesserte Reliabilität bei der Testdurchführung mit dem Rangreihenverfahren im Vergleich zu einem herkömmlichen Multiple-Choice-Test. Kritisch ist anzumerken, dass Diamond (1975) keine admissible Auswertefunktion verwendete. Testteilnehmer können bei diesem Verfahren ihre Punktsumme auch dann maximieren, wenn sie ihr Wissen verfälscht wiedergeben. Bei völligem Nichtwissen bezüglich aller Antwortoptionen entspricht diese Auswertung deshalb einem herkömmlichen Multiple-Choice-Verfahren (Curlette, 1978). Anderson (1982) beurteilte das Rangreihenverfahren generell als unzureichend. Die Teilnehmer würden in der Regel nicht alle Antwortoptionen bewerten und in eine Reihenfolge bringen, sondern maximal zwei oder drei.

5.2 Verbale Skalen

Verbale Skalen bilden Antwortsicherheit in Form von Beschreibungen wie beispielsweise „nicht sicher“, „schwach sicher“, „relativ sicher“ und „sehr sicher“ ab. Bokhorst (1986) verwendete eine bipolare verbale Skala zur Erhebung von Konfidenzurteilen. Er untersuchte 497 Teilnehmer mithilfe eines Tests, der aus 40 Items mit vier Antwortoptionen bestand. Die Testteilnehmer wählten zuerst die ihrer Meinung nach richtige Antwort aus und gaben dann auf einer verbalen Skala, die aus den beiden Werten „sicher“ und „unsicher“ bestand, an, wie sicher sie sich für die jeweilige Antwortoption entschieden hatten. Die Punktauszahlung erfolgte nach folgendem Verfahren: Vier Punkte erhielten die Testteilnehmer, wenn sie die richtige Antwortoption „sicher“ gewählt hatten. Einen Punkt erhielten sie, wenn sie zwar die richtige Antwort gewählt hatten, diese Wahl aber als „unsicher“ bewertet hatten. Bei einer „sicher“ gewählten falschen Antwort wurden den Teilnehmern vier Punkte abgezogen, bei einer „unsicheren“ falschen Antwort verloren sie einen Punkt. Entschieden sich die Teilnehmer für keine Antwortoption, so erhielten sie null Punkte. Als Multiple-Choice-Kontrollbedingung betrachtete Bokhorst (1986) die ausgewählten Antwortoptionen. Er konnte eine verbesserte Reliabilität der Bedingung mit den verbalen Konfidenzurteilen gegenüber der Multiple-Choice-Bedingung zeigen. Eine höhere Validität fand er hingegen nicht. Bokhorst (1986) verwendete ebenfalls keine admissible Auswertefunktion nach Shuford, Albert und

Massengill (1966). Die Auswertung hat aber gegenüber einer herkömmlichen dichotomen Auszahlung von null Punkten für eine falsche Antwort und einem Punkt für eine richtige Antwort den Vorteil, dass sie Raten nicht belohnt. Generell haben verbale Skalen den Nachteil, dass sie vage sind, da jeder Teilnehmer etwas anderes unter dem jeweiligen Begriff verstehen kann (Leclerq, 1983).

5.3 Punktesysteme

Bei Punktesystemen ordnen Testteilnehmer ihre Antwortsicherheiten den einzelnen Antwortoptionen zu, indem sie eine vorgegebene Anzahl von Punkten auf die Optionen verteilen. Michael (1968) verwendete ein Punktesystem in einem Experiment mit 432 Testteilnehmern. Das Testmaterial bestand aus 35 Items mit vier Antwortoptionen. Die Teilnehmer führten den Test zuerst als herkömmlichen Multiple-Choice-Test durch. Dazu verwendeten sie einen roten Stift, um zunächst die richtige Antwortoption zu markieren. Nach Abschluss des Tests mussten die Teilnehmer den roten gegen einen schwarzen Stift tauschen. Mit diesem verteilten sie dann jeweils zehn Punkte auf die vier Antwortoptionen, um ihre Antwortsicherheit bezüglich aller Optionen anzugeben. Die beiden unterschiedlichen Stiftfarben sollten dabei sicherstellen, dass sie nachträglich nicht mehr die anfangs getroffene Wahl der am wahrscheinlichsten richtigen Antwortoption ändern konnten. Die Punkte, die die Testteilnehmer der richtigen Antwortoption zugeordnet hatten, erhielten sie als Punkte ausgezahlt. Die mit dem roten Stift markierten Antwortoptionen wertete Michael (1968) als Multiple-Choice-Bedingung unter einer herkömmlichen dichotomen Auswertung aus. Zusätzlich führte er eine Ratekorrektur durch. Diese beiden Multiple-Choice-Bedingungen verglich Michael (1968) mit der Bedingung, in der die Verteilung der Punkte vorgenommen wurde. Dabei konnte er eine deutlich verbesserte Reliabilität der Bedingung mit dem Punktesystem gegenüber der herkömmlichen Multiple-Choice-Bedingung und der Auswertung unter Ratekorrektur zeigen. Michael (1968) beurteilte die Verteilung von Konfidenz-punkten als ein Antwortverfahren, das in Klassenräumen eingesetzt werden könne und dazu geeignet sei, ein hohes Maß an Informationen über den Wissensstand eines Testteilnehmers zu erheben. Kritisch ist zu sehen, dass Michael (1968) die Testteilnehmer zuerst aufforderte, nur eine Antwortoption auszuwählen, d.h. bei

Nicht- oder Teilwissen zu raten, und erst in einem zweiten Durchgang die Punkteverteilung vorzunehmen. Es ist zu vermuten, dass die Teilnehmer die höchste Punktzahl der bereits gewählten Antwortoption zuordneten und nicht mehr alle Antwortoptionen sorgfältig beurteilten. Zudem wurden die Teilnehmer durch die Verteilung von zehn Punkten auf vier Antwortoptionen gezwungen, zumindest eine Antwortoption als die am wahrscheinlichsten richtige auszuwählen. Die Verteilung einer Anzahl von Punkten, die nicht gleichmäßig auf die vorhandenen Antwortoptionen aufgeteilt werden kann, ist jedoch ungeeignet, um Antwortsicherheit zu erheben. Die Teilnehmer können dann völliges Nichtwissen nicht angeben und haben keine andere Möglichkeit, als eine Antwortoption zu favorisieren (Anderson, 1982). Auch die Auszahlung der Konfidenzpunkte als Testpunkte ist nicht dazu geeignet, Teilnehmer zu einer unverfälschten Reproduktion ihrer tatsächlichen subjektiven Antwortsicherheiten zu inzentivieren. Denn unter dieser Auswertung ist es die beste Strategie, alle Punkte auf die plausibelste Antwortoptionen zu setzen, wenn ein Teilnehmer das Ziel hat, seine Gesamtpunktsumme zu maximieren.

Rippey (1970) verwendete Items mit drei Antwortoptionen und ließ die Testteilnehmer neun Punkte verteilen. Damit hatten sie die Möglichkeit, völliges Nichtwissen einzuräumen, indem sie jeder Antwortoption drei Punkte zuordneten. Rippey (1970) verglich in einem Experiment vier verschiedene Auswertefunktionen: eine lineare, eine sphärische, eine euklidische und eine logarithmische Funktion nach Shuford, Albert und Massengill (1966) sowie eine herkömmliche dichotome Multiple-Choice-Auswertung. Die Teilnehmer erhielten in allen Versuchsbedingungen dieselbe Information über die Auswertung. Es wurde darauf hingewiesen, dass die Teilnehmer nur dann ihre Gesamtpunktsumme maximieren könnten, wenn sie ihre Antwortsicherheiten unverfälscht berichteten. Sie wurden aber nicht über die genaue Auszahlung der Punkte mit der jeweiligen Auswertefunktion informiert. Der Vergleich der Auswertefunktionen erfolgte auf der Basis derselben Rohdaten, d.h. der Punkte, die den Antwortoptionen zugeordnet worden waren. Rippey (1970) errechnete für jede der Funktionen die Punktauszahlungen je Teilnehmer. Bei einer linearen Auswertung zeigte der Test die höchste Reliabilität und bei einer dichotomen Multiple-Choice-Auswertung die niedrigste. Die Reliabilitäten, ermittelt beim Einsatz einer logarithmischen, sphärischen und euklidischen Auswerte-

funktion, waren nur wenig besser als die der Multiple-Choice-Auswertung. Rippey (1970) sah die Ursache für die schlechteren Reliabilitäten bei einer Auswertung mithilfe einer logarithmischen, sphärischen oder euklidischen Funktion im Vergleich zur linearen in einem geringen Verständnis der Art und Weise der Auswertung durch die Teilnehmer. Die lineare Auszahlung von Punkten, die identisch sind mit den prozentualen Antwortsicherheiten in die richtige Antwort, sei am intuitivsten. Kritisch ist zu sehen, dass die Teilnehmer während des Tests nicht über die tatsächliche Punktauszahlung informiert wurden. So kann keine Aussage darüber getroffen werden, wie sie sich unter einer solchen Information verhalten hätten. Rippey und Voytovich (1983) kamen selbst zu dem Schluss, dass es unerwünschte Konsequenzen auf die Reliabilität haben könnte, wenn die Teilnehmer nicht während eines Tests über die genaue Punktauszahlung informiert würden.

Kansup und Hakstian (1975) verwendeten ebenfalls ein Punktesystem, um verschiedene Auswertefunktionen, darunter eine lineare und eine logarithmische Funktion nach Shuford, Albert und Massengill (1966), zu untersuchen. Sie testeten 348 Studenten und verwendeten Items mit vier oder fünf Antwortoptionen. Die Studenten verteilten zehn Punkte auf die Antwortoptionen, und auch in diesem Experiment wurden die Auswertefunktionen nicht explizit erläutert. Die Testteilnehmer wurden instruiert, dass sie ihre Punktsumme nur dann maximieren könnten, wenn sie ihre Antwortsicherheiten unverfälscht berichteten. Für die von den Teilnehmern angegebenen Punkteverteilungen berechneten Kansup und Hakstian (1975) die Punktauszahlungen mithilfe verschiedener Auswertefunktionen. Auf der Basis der ausgezahlten Punkte je Item kalkulierten sie die Reliabilität und die Validität und fanden keine signifikanten Unterschiede zwischen den Auswertebedingungen. Kansup und Hakstian (1975) folgerten deshalb, dass die Wahl der Auswertefunktion relativ unwichtig sei und sahen durch ihr Experiment die Ergebnisse des Experiments von Rippey (1970) gestützt. Auch das Experiment von Kansup und Hakstian (1975) weist den Schwachpunkt auf, dass die Teilnehmer während der Testdurchführung nicht über die Punktauszahlung informiert wurden.

5.4 Geometrische Figuren

Zur Erhebung von Antwortsicherheit können auch geometrische Figuren eingesetzt werden. Toda (1963) verwendete beispielsweise zwei Quadrate, um den Testteilnehmern die zu erwartende Punktauszahlung zu visualisieren. Dies geschah, indem die Größe der Quadrate in Abhängigkeit von der Antwortsicherheit, die ein Teilnehmer angab, verändert wurde. Dieses Instrument verwendete Toda (1963) mit einer Auswertung anhand einer quadratischen Funktion. Um die Punktauszahlung durch eine logarithmische Funktion zu visualisieren, verwendete er eine Linie. Toda (1963) führte mit beiden Instrumenten Pilotstudien durch. Er berichtete aber keine konkreten Ergebnisse, sondern nur die Beobachtungen, dass Testteilnehmer unter Einsatz der Quadrate und der quadratischen Funktion eher zu hohen Sicherheitsurteilen neigten, während sie unter Verwendung der Linie und der logarithmischen Funktion eher vorsichtiger waren. Toda (1963) kommentierte seine Beobachtungen, dass er nicht wisse, ob die Ursache für dieses unterschiedliche Antwortverhalten in den verschiedenen Auswertefunktionen oder in den verschiedenen Instrumenten zur Erhebung von Antwortsicherheit zu suchen sei.

Bei Items mit drei Antwortoptionen kann zur Erhebung von Antwortsicherheit auch eine Dreiecksfigur eingesetzt werden. Jeder Ecke des Dreiecks wird dabei eine der Antwortoptionen zugeordnet. Jeder Punkt der Dreiecksfläche entspricht einer bestimmten simultanen Verteilung von Antwortsicherheit auf die drei Antwortoptionen (De Finetti, 1965; Shuford & Brown, 1975). Rippey (1979) verwendete ein solches Dreieck als Antwortinstrument in einem Experiment. Alle bisher beschriebenen Experimente wurden in Papier-und-Bleistift-Form durchgeführt. Rippey (1979) testete die Teilnehmer mithilfe eines Computerprogramms an einem PLATO-Terminal. Auf dem Bildschirm war, neben der Frage und den Antwortoptionen, das Dreieck dargestellt. Um ihre Antwort abzugeben, bewegten die Testteilnehmer durch Tastatureingaben einen *Cursor* im Dreieck auf die gewünschte Position. Die Auszahlung der Punkte erfolgte mithilfe einer logarithmischen Auswertefunktion nach Shuford, Albert und Massengill (1966). Die zu erwartenden Punktauszahlungen wurden, den aktuell eingestellten prozentualen Antwortsicherheiten entsprechend, an den Ecken des Dreiecks angezeigt. Die Teilnehmer wurden so über die mögliche Konsequenz, hohe negative Punktauszahlungen zu erhalten, informiert, bevor sie ihre Entscheidung trafen. Nach

jedem Item erhielten sie ein *Feedback*, das aus der richtigen Antwortoption, der Punktauszahlung sowie der Gesamtpunktsumme, dargestellt als Diagramm, bestand. An diesem Experiment nahmen 36 Medizinstudenten teil. Rippey (1979) korrigierte die Antwortsicherheiten der Testteilnehmer mithilfe eines individuellen Realismusindex nach Brown und Shuford (1973) und errechnete die Punktauszahlungen auf der Basis der korrigierten Antwortsicherheiten erneut. Rippey (1979) konnte zeigen, dass die Reliabilität und die Validität des Tests, ermittelt für die korrigierten Punktauszahlungen, höher waren als die, die auf der Basis der nicht korrigierten ermittelt wurden. Der Unterschied war jedoch nicht signifikant. Rippey (1979) sah in diesem Ergebnis trotzdem einen Anreiz, dieses Verfahren weiterhin einzusetzen. Ein Vergleich mit einem herkömmlichen Multiple-Choice-Verfahren erfolgte nicht.

Rippey und Voytovich (1983) untersuchten mithilfe eines Computerprogramms das Dreieck von Rippey (1979) und ein weiteres Antwortdreieck. Das zweite Dreieck war in viele kleine Dreiecke unterteilt. Jedes dieser kleinen Dreiecke war beschriftet mit den prozentualen Antwortsicherheiten, die seine Auswahl bedeutete und der zu erwartenden Punktauszahlung. Ein Teilnehmer beantwortete ein Item, indem er die Nummer eines der kleinen Dreiecke in den Computer eingab. Die Auswertung erfolgte durch eine logarithmische Funktion nach Shuford, Albert und Massengill (1966). Rippey und Voytovich (1983) führten mit diesen beiden Antwortdreiecken drei Experimente mit je 37, 76 und 83 Studenten durch. Sie zeigten eine Verbesserung der Reliabilität und der Validität des Tests aufgrund der Korrektur der Antwortsicherheiten anhand des individuellen Realismusindex. Einen Vergleich der beiden Antwortdreiecke berichteten Rippey und Voytovich (1983) nicht. Auch ein Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren erfolgte nicht.

Bruno (1986) adaptierte ebenfalls ein Dreieck an eine Testdurchführung mit einem Computer. Er unterteilte jede Seitenlinie des Dreiecks durch drei Trennstriche in vier Abschnitte. Jede Trennlinie, jede Ecke und die Mitte der Dreiecksfläche kennzeichnete er fortlaufend mit Buchstaben, so dass 13 Auswahlmöglichkeiten entstanden. Da nur die Außenlinie des Dreiecks verwendet wurde, konnten sich die Testteilnehmer nur zwischen zwei Antwortoptionen entscheiden, wenn sie bezüglich der richtigen Antwort unsicher waren, und mussten eine der drei Optionen völlig ausschließen. Die Kennzeichnung der Mitte des Dreiecks ermöglichte es ihnen, auch

völliges Nichtwissen einzuräumen. Die Punktauszahlung erfolgte anhand einer logarithmischen Auswertefunktion nach Shuford, Albert und Massengill (1966). Das Dreieck wurde auf einem Computermonitor über dem Itemtext angezeigt. Die Teilnehmer beantworteten ein Item, indem sie den Buchstaben, der ihrer Antwortsicherheit entsprach, auf einer Tastatur eingaben. Bruno (1986) analysierte das Antwortverhalten der Testteilnehmer und fand, dass das Antwortverfahren besonders für Teilnehmer mit Teilwissen von Vorteil sei, da sie für die unverfälschte Reproduktion ihres Wissens durch Punktauszahlungen belohnt würden. Er vermutete, dass mit diesem Antwortverfahren weniger Items benötigt würden, um den Wissensstand eines Testteilnehmers zu messen, als mit Multiple-Choice. Testgütekriterien des Tests mit Multipler Evaluation im Vergleich zu Multiple-Choice berichtete Bruno (1986) jedoch nicht.

Abedi und Bruno (1989) adaptierten das Dreieck von Bruno (1986) auch an eine Durchführung als Papier-und-Bleistift-Version. Dieses Dreieck konnte optisch *gescannt* werden, was eine einfachere Auswertung ermöglichte. Abedi und Bruno (1989) führten Tests zu verschiedenen Wissensgebieten an insgesamt 456 Studenten durch. Dabei verglichen sie das Antwortverfahren Multiple Evaluation mit einer logarithmischen Auswertung nach Shuford, Albert und Massengill (1966) und der Dreiecksfigur als Antwortinstrument mit einem herkömmlichen Multiple-Choice-Verfahren. Abedi und Bruno (1989) zeigten eine deutlich verbesserte Retest-Reliabilität des Tests mit Multipler Evaluation im Vergleich zu Multiple-Choice. Eine höhere Reliabilität fanden sie vor allem bei schwierigen Items. Sie führten dieses Ergebnis darauf zurück, dass die Varianz der beobachtbaren Testwerte aufgrund einer verfälschten Wiedergabe der tatsächlichen subjektiven Antwortsicherheiten reduziert werden konnte. Da diese Varianz bei schwierigen Items in der Regel am höchsten ist, zeigte sich hier auch die deutlichste Verbesserung. In einem weiteren Experiment, in dem sie dieses Verfahren ebenfalls einsetzten, zeigten Abedi und Bruno (1993) eine Verbesserung der Validität eines Tests mit Multipler Evaluation im Vergleich zu einem Multiple-Choice-Test.

5.5 Prozentskalen

Prozentskalen sind eine weitere Möglichkeit, Antwortsicherheit zu erfassen. Mit solchen Skalen kann Antwortsicherheit sehr differenziert, z.B. in Schritten von einem Prozent, erhoben werden. Basierend auf einer Prozentskala entwickelte Shuford (1969) das Instrument „SCoRule™“, welches aus mehreren Scheiben bestand. Indem er diese Scheiben drehte, konnte ein Testteilnehmer die gewünschte prozentuale Antwortsicherheit einstellen und für diese die zu erwartende Punktauszahlung ablesen. So informierte Shuford (1969) einen Teilnehmer während der Bearbeitung eines Items über die Auszahlung. Die Antwortmöglichkeiten waren durch Buchstaben kodiert. Den Buchstaben, der seiner Antwort entsprach, notierte ein Teilnehmer auf seinem Antwortbogen. Shuford (1971) testete unter Einsatz dieses Instruments Teilnehmer eines militärischen Ausbildungsseminars. Er fand eine Verbesserung der Validität des Tests mit Multipler Evaluation im Vergleich zu einem herkömmlichen Multiple-Choice-Test. Ebel (1968) kritisierte an dem von Shuford (1969) entwickelten Testinstrument, dass sich die Durchführungszeit eines Tests damit verdopple und die Kosten pro Test relativ hoch seien.

Hambleton, Roberts und Traub (1970) erhoben Antwortsicherheit in einem Experiment mit einer Prozentskala in Form einer Papier-und-Bleistift-Version. Das Testmaterial bestand aus Items mit fünf Antwortoptionen. Die Skala war in Schritte von je fünf Prozent unterteilt. Unter dieser Skala befanden sich für jede Antwortoption 20 Kästchen waagerecht nebeneinander. Die Teilnehmer beantworteten ein Item, indem sie ihre Antwortsicherheit als waagerechte Linie entlang der Kästchen einzeichneten. Durch einen senkrechten Strich sollten sie die Linie zu den Kästchen einer weiteren Antwortoption ziehen und dort die Linie waagerecht nach rechts weiterzeichnen. Die Auswertung erfolgte mithilfe einer logarithmischen Funktion nach Shuford, Albert und Massengill (1966). Hambleton, Roberts und Traub (1970) fanden eine verbesserte Validität, aber eine schlechtere *Split-Half-Reliabilität* bei einer Erhebung von Antwortsicherheit mit dieser Skala im Vergleich zu der Multiple-Choice-Bedingung. Sie kommentierten ihre Ergebnisse aber aufgrund zu geringer Teilnehmerzahlen (N=71) als nicht signifikant. Außerdem seien die eingesetzten Items mit einer mittleren Lösungswahrscheinlichkeit von 75% zu einfach gewesen. Diese Kritik ist berechtigt, da anzunehmen ist, dass die Varianz der beobachtbaren Testwerte bei einfachen Items wesentlich geringer ist als bei

schwierigen. Deshalb ist eine Verbesserung der Testgüte, aufgrund einer Reduzierung dieser Varianz, besonders bei schwierigen Items zu erwarten.

Romberg, Shepler und Wilson (1970) setzten ebenfalls das Antwortverfahren Multiple Evaluation in Form einer Papier-und-Bleistift-Version ein. Der Test bestand aus zehn Mathematikaufgaben mit je fünf Antwortoptionen. Die Testteilnehmer notierten ihre Antwortsicherheiten in Wahrscheinlichkeiten zwischen null und eins auf einem Antwortblatt. Dabei sollten sie Schritte von 0,1 verwenden. Romberg, Shepler und Wilson (1970) führten drei Experimente durch. An ihrem ersten Experiment nahmen pro Versuchsgruppe 32 Studenten teil. Die Kontrollbedingung war ein herkömmliches Multiple-Choice-Verfahren. Die Studenten erhielten vor dem Test eine kurze Instruktion darüber, wie sie ihre Antwortsicherheit angeben sollten und wurden zudem darüber informiert, dass sie nur dann ihre Gesamtpunktsumme maximieren könnten, wenn sie ihre Antwortsicherheit vollkommen unverfälscht wiedergäben. Die Bedingungen mit Multipler Evaluation und der Auswertung mit einer logarithmischen Funktion zeigte eine schlechtere Reliabilität gegenüber der Multiple-Choice-Bedingung. Romberg, Shepler und Wilson (1970) vermuteten, dass diese Ergebnisse darauf zurückzuführen seien, dass die Testitems nicht die erwarteten Itemschwierigkeiten gezeigt hatten, sondern wesentlich einfacher waren. In einem zweiten Experiment verwendeten sie 17 Mathematikaufgaben und intensivierten die Instruktion der Testteilnehmer durch Übungen. Auch in diesem Experiment konnten Romberg, Shepler und Wilson (1970) mit Multipler Evaluation und einer logarithmischen Auswertung keine Verbesserung der Reliabilität gegenüber der Multiple-Choice-Bedingung zeigen. Sie vermuteten, dass schwierige Mathematikaufgaben kein geeignetes Testmaterial seien. Außerdem nahmen sie an, dass die Studenten noch intensiver über die logarithmische Auswertung hätten aufgeklärt werden müssen, um ihr Verhalten im Test zu beeinflussen. In einem dritten Experiment verwendeten Romberg, Shepler und Wilson (1970) 30 Items eines Intelligenztests. Am Tag vor dem Test wurden die Testteilnehmer intensiv in das Antwortverfahren eingeführt. Zusätzlich wurde die Instruktion unmittelbar vor dem Test wiederholt. Die Multiple-Choice-Kontrollbedingung bestand aus 67, die Bedingung mit Multipler Evaluation aus 58 Teilnehmern. Auch die Ergebnisse dieses Experiments zeigten eine verschlechterte Reliabilität mit Multipler Evaluation und einer logarithmischen Auswertung gegenüber dem herkömmlichen Multiple-Choice-

Verfahren. Romberg, Shepler und Wilson (1970) folgerten aus den Ergebnissen ihrer Experimente, dass das Problem, den Wissensstand eines Teilnehmers bei schwierigen Items unverfälscht zu messen, noch nicht gelöst sei.

Koehler (1971) verglich die Erhebung von Antwortsicherheit mithilfe einer Wahrscheinlichkeitsskala mit der Erhebung mit einem Punktesystem, bei dem die Teilnehmer zehn Punkte auf die Antwortoptionen verteilen sollten. Beide Antwortverfahren setzte er in Form einer Papier-und-Bleistift-Version ein. Als Testmaterial verwendete er Items mit fünf Antwortoptionen. Die Auszahlung der Punkte erfolgte anhand einer quadratischen sowie einer logarithmischen Funktion nach Shuford, Albert und Massengill (1966). Koehler (1971) untersuchte 535 Testteilnehmer, verteilt auf drei Versuchsbedingungen. Die Reliabilität berechnete er auf der Basis von jeweils zehn Items für drei verschiedene Fachgebiete. Koehler (1971) konnte für keine der Bedingungen mit Multipler Evaluation eine verbesserte Reliabilität oder Validität gegenüber dem herkömmlichen Multiple-Choice-Verfahren zeigen. Er führte die Ergebnisse darauf zurück, dass zehn Items nicht genügten, um signifikante Unterschiede bezüglich der Reliabilität und der Validität eines Tests zu messen. Koehler (1971) schloss aber trotzdem, dass das herkömmliche Multiple-Choice-Verfahren einem Test mit Multipler Evaluation vorzuziehen sei, da die Testzeit kürzer, die Administration einfacher und kein Training der Testteilnehmer erforderlich sei.

Rippey (1968a) entwickelte ein Computerprogramm zur Erhebung von Antwortsicherheit. Das genaue Antwortverfahren beschrieb er jedoch nicht. Unter Einsatz dieses Programms führte er verschiedene Experimente durch, um die These von Shuford, Albert und Massengill (1966) zu prüfen, dass bei der Durchführung eines Tests mit Multipler Evaluation mit der Auswertung durch eine admissible logarithmische Auswertefunktion eine Verbesserung der Reliabilität im Vergleich zu einem herkömmlichen Multiple-Choice-Test zu beobachten sei. Rippey (1968b) fand ein inkonsistentes Muster, in dem sich sowohl Verbesserungen als auch Verschlechterungen der Reliabilität des Tests mit Multipler Evaluation im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren zeigten. Darüber hinaus beobachtete Rippey (1968b) eine gute Akzeptanz der Prozedur durch die Teilnehmer, aber eine deutlich verlängerte Durchführungszeit.

Koele, De Boo und Verschure (1987) führten ein Experiment mit einem Test mit Multipler Evaluation in einer Papier-und-Bleistift-Version durch. Die Teilnehmer verteilten 100% Antwortsicherheit auf vier Antwortoptionen, indem sie ihre jeweilige prozentuale Antwortsicherheit vor die Antwortoptionen schrieben. 125 Studenten nahmen an diesem Experiment teil. Der verwendete Test bestand aus 50 Items. Koele, De Boo und Verschure (1987) untersuchten in den Bedingungen mit Multipler Evaluation zwei verschiedene Auswertefunktionen im Vergleich zu einem herkömmlichen Multiple-Choice-Test. Eine davon war eine lineare Auswertefunktion, die die prozentuale Antwortsicherheit, die ein Testteilnehmer der richtigen Antwortoption zugeordnet hatte, in Punkten auszahlte. Die zweite Funktion war die Brier-Regel. Beide Auswertebedingungen des Tests mit Multipler Evaluation zeigten gegenüber der Bedingung mit Multiple-Choice eine verbesserte Reliabilität, wobei die lineare Auswertung die höchste Verbesserung bewirkte. Insgesamt hatte dieser Test einen hohen Schwierigkeitsgrad. In der Bedingung mit Multiple-Choice betrug die Anzahl richtig gelöster Items nur 25%.

Dirkzwager (1993) verwendete Schieberegler, implementiert in einer Computeranwendung, um Antwortsicherheiten zu erfassen. Jeder Antwortoption war ein Schieberegler zugeordnet, der auf ganzzahlige Werte von 0% bis 100% eingestellt werden konnte. In der Summe ergaben die Werte aller Schieberegler 100%. Rechts neben den Schiebereglern wurden als *Feedforward* die zu erwartenden Punktauszahlungen angezeigt. Die Auswertung erfolgte durch eine logarithmische Funktion. Direkt im Anschluss an die Bearbeitung eines Items erhielten die Testteilnehmer ein *Feedback*, das aus der richtigen Antwortoption und den ausgezahlten Punkten bestand. Dirkzwager (1996) führte mit dieser Testapplikation ein Experiment mit 47 im Durchschnitt elfjährigen Schülern durch. Anhand eines verbalen Intelligenztests konnte er zeigen, dass auch Kinder in der Lage sind, das Prinzip der Multiplen Evaluation zu verstehen und ihr Wissen damit zu reproduzieren. Die Reliabilität oder die Validität des Tests dieses Experiments berichtete Dirkzwager (1996) nicht.

Holmes (2002) setzte das Antwortverfahren Multiple Evaluation zur Wissensüberprüfung bei 60 Studenten eines Seminars für Computertechnik ein. Er führte zehn Tests durch und verglich eine computerunterstützte Anwendung mit einer Papier-und-Bleistift-Version. Für die computerunterstützte Durchführung verwen-

dete er die Testapplikation von Dirkzwager (1993). In dieser Bedingung bearbeiteten die Teilnehmer fünf Items mit je vier Antwortoptionen. Bei der Papier-und-Bleistift-Version hatten die Testteilnehmer neun Prozentkategorien zur Verfügung, um ihre Antwortsicherheiten zu berichten. In dieser Versuchsbedingung bearbeiteten die Teilnehmer zehn Items mit je zwei Antwortoptionen. Holmes (2002) verwendete eine logarithmische Auswertefunktion nach Dirkzwager (2003). Dabei wählte er den Toleranzfaktor T=4. Damit sah die Auswertung bei nur geringen Antwortsicherheiten in die richtige Antwort hohe Strafzahlungen vor. Bei der Papier-und-Bleistift-Version mussten die Testteilnehmer die zu erwartende Punktauszahlung aus einer Tabelle ablesen. In der Bedingung, die mit der Computerapplikation durchgeführt wurde, erhielten die Testteilnehmer die Information über die Auszahlung durch die Applikation. Die Teilnehmer der Bedingung mit der Papier-und-Bleistift-Version erhielten eine detailierte Rückmeldung, nachdem der Test ausgewertet worden war. Holmes (2002) schloss auf der Basis der Ergebnisse, dass das Antwortverfahren Multiple Evaluation im Prüfungsalltag einsetzbar sei. Er fand eine verbesserte Reliabilität in beiden Bedingungen mit Multipler Evaluation gegenüber der Bedingung mit Multiple-Choice. Die Daten der Bedingung mit Multiple-Choice erhob Holmes (2002) jedoch nicht an einer Versuchsgruppe, sondern er errechnete sie aus den mit dem Antwortverfahren Multiple Evaluation erhobenen Daten.

5.6 Zusammenfassung und Schlussfolgerungen

Die dargelegten empirischen Befunde konnten die Annahme, dass durch den Einsatz des Antwortverfahrens Multiple Evaluation mit einer logarithmischen Auswertung die Varianz der beobachtbaren Testwerte eines Tests reduziert und damit die psychometrische Güte eines Tests verbessert werden kann, nicht eindeutig belegen. Nur teilweise konnten Verbesserungen der Reliabilität und der Validität im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren gezeigt werden. Eine Ursache dafür ist möglicherweise in der praktischen Umsetzung des Verfahrens zu suchen (Rippey & Voytovich, 1985). Die vorgestellten unterschiedlichen Versuche, ein geeignetes Antwortinstrument für die Erhebung von Antwortsicherheit zu entwickeln, lassen die Herausforderungen bei der Implementierung erahnen. Die Anforderungen an die Testdurchführung können in Form einer Papier-und-Bleistift-

Version kaum umgesetzt werden. So sollte das Antwortinstrument beispielsweise gewährleisten, dass die Antwortsicherheiten eines Teilnehmers in der Summe immer 100% ergeben. Nur so kann eine unvollständige Angabe von Antwortsicherheit verhindert werden. Eine weitere Herausforderung ist eine Information des Testteilnehmers über die Punktauszahlung, während er ein Item bearbeitet. Rippey und Voytovich (1983, S. 88) sahen die Lösung dieser Probleme in der Verwendung eines Computers: *"Difficulty of implementation, however, has been one of the inherent problems of confidence testing from its inception, and the use of computers has been the best solution."* Auch Shuford (1993) sah nach erfolglosen Versuchen keine Möglichkeit, das Verfahren ohne den Einsatz eines Computers zu verwenden (Shuford, 1993, S. 1): *"So I undertook a long pilgrimage in search of a way to implement this procedure without using a computer. I failed- …"*

Ein Kritikpunkt an einigen Experimenten ist, dass die Auszahlung der Punkte nicht anhand einer admissiblen Auswertefunktion erfolgte. Nur mit einer Auszahlung durch eine admissible Auswertefunktion wird sichergestellt, dass ein Teilnehmer seine Gesamtpunktsumme nur dann maximieren kann, wenn er seine Antwortsicherheiten völlig unverfälscht berichtet (Dirkzwager, 2003; Shuford, Albert & Massengill, 1966). Erfolgte die Punktauszahlung im Experiment durch eine logarithmische Auswertefunktion, so wurden die Testteilnehmer häufig während der Bearbeitung eines Items nicht, wie beispielsweise von Dirkzwager (1993) gefordert, über die zu erwartende Punktauszahlung informiert. Besonders kritisch ist in diesem Zusammenhang auch zu sehen, wenn die Teilnehmer überhaupt nicht über die der Auszahlung zugrunde liegende Auswertefunktion informiert wurden. In den Experimenten von Kansup und Hakstian (1975) oder Rippey (1970) wurden beispielsweise die Ergebnisse für verschiedene Auswertefunktionen auf der Basis derselben Antwortsicherheiten nur errechnet. So kann keine Aussage darüber getroffen werden, wie die Testteilnehmer ihre Antwortsicherheiten berichtet hätten, wenn sie im Test über die Punktauszahlung informiert worden wären.

Den genannten Kritikpunkten ist hinzuzufügen, dass im Design mancher Experimente, beispielsweise Holmes (2002), Michael (1968) oder Shuford (1969), eine echte Kontrolle fehlte (Echternacht, 1972). Die Daten der Kontrollbedingung mit Multiple-Choice wurden nicht im Experiment gemessen, sondern aus den Antwortsicherheiten der Bedingungen mit Multipler Evaluation ermittelt. Dazu

wurde diejenige Antwortoption als Antwort gewertet, der die höchste Antwortsicherheit zugeordnet worden war. Die Erzeugung einer Multiple-Choice-Bedingung aus den mit dem Antwortverfahren Multiple Evaluation erhobenen Antwortsicherheiten führt jedoch nicht zu eindeutigen Ergebnissen. Denn in den Fällen, in denen ein Teilnehmer keine der Antwortoptionen bevorzugt hat, also in denen er beispielsweise bei Items mit zwei Antwortoptionen eine Sicherheit von 50% in jede Option berichtet hat, muss der Testleiter eine Entscheidung treffen, ob und wie er Punkte auszahlt. Dies wird anhand der folgenden Beispielrechnung verdeutlicht. Grundlage der Berechnung bilden die Antwortsicherheiten von 514 Testteilnehmern einer Bedingung mit Multipler Evaluation eines Experiments dieser Arbeit. Betrachtet werden die Antwortsicherheiten in die richtige Antwort von 36 Items mit je zwei Antwortoptionen. Zunächst wurde für eine Antwortsicherheit in die richtige Antwort größer als 50% ein Punkt ausgezahlt und für eine Antwortsicherheit kleiner als 50% null Punkte. Drei unterschiedliche Berechnungsmöglichkeiten ergaben sich für den Fall, in denen ein Testteilnehmer jeder Antwortoption 50% Antwortsicherheit zugeordnet hatte. Eine Möglichkeit war, dem Teilnehmer in diesem Fall einen Punkt auszuzahlen. Damit wurde ein Cronbach-α von 0,70 errechnet. Wurden stattdessen null Punkte vergeben, so errechnete sich ein Cronbach-α-Koeffizient von 0,81. Bei einer zufälligen Punktauszahlung von null oder einem Punkt betrug Cronbach-α 0,74. Die Prüfung

Tabelle 5-1

Prüfung auf signifikante Unterschiede zwischen den errechneten Cronbach-α-Koeffizienten.

Auszahlung bei jeweils 50% Antwortsicherheit	α	Auszahlung bei jeweils 50% Antwortsicherheit	α	X^2	p	Anzahl Items	N
Auszahlung von 1 Punkt	0,70	zufällige Auszahlung von 0 oder 1 Punkt	0,74	3,2	,07	36	514
Auszahlung von 1 Punkt	0,70	Auszahlung von 0 Punkten	0,81	33,8	<,001**	36	514
zufällige Auszahlung von 0 oder 1 Punkt	0,74	Auszahlung von 0 Punkten	0,81	16,2	<,001**	36	514

Die Prüfung auf signifikante Unterschiede erfolgte durch das Programm alphatst.exe (Lautenschlager, 1989) nach der Methode von Feldt, Woodruff und Salih (1987).
**=signifikant auf einem Signifikanzniveau von 0,1%.

auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten, dargestellt in Tabelle 5-1, zeigt, dass die Auszahlung von einem Punkt zu einem um 0,11

signifikant geringeren α-Koeffizienten führt als die Auszahlung von null Punkten. Auch die zufällige Auszahlung von null oder einem Punkt, verglichen mit der Auszahlung von immer null Punkten, bewirkt einen signifikant um 0,07 geringeren Cronbach-α-Koeffizienten. Die Erzeugung einer Multiple-Choice-Kontrollbedingung aus erhobenen Antwortsicherheiten stellt damit also keine aussagekräftige Kontrolle dar.

In den meisten der vorgestellten Experimente wurde die Schwierigkeit der Items nicht kontrolliert. Es ist jedoch anzunehmen, dass sich der Vorteil des Antwortverfahrens Multiple Evaluation, nämlich die Varianz der beobachtbaren Testwerte eines Tests zu verringern, besonders bei schwierigen Items zeigt. Abedi und Bruno (1989) konnten diesen Einfluss der Schwierigkeit zeigen. Sie fanden eine Verbesserung der Reliabilität der Multiplen Evaluation im Vergleich zu Multiple-Choice besonders bei schwierigen Items.

Die Kritik an den Experimenten muss jedoch vor dem Hintergrund der technischen Möglichkeiten gesehen werden. Die meisten Experimente wurden durchgeführt, als die Computertechnik noch nicht weit genug entwickelt war und Computer kaum zur Verfügung standen. Mithilfe der heutigen technischen Möglichkeiten stellt es hingegen kein Problem mehr dar, den Anforderungen an einen Test mit Multipler Evaluation gerecht zu werden. Computer stehen in den meisten Bildungseinrichtungen und Haushalten zur Verfügung. Die Nutzung eines Computers bietet zudem auch die Möglichkeit, die Experimente webbasiert durchzuführen. So kann ohne enormen Aufwand eine Stichprobe gemessen werden, die genügend groß ist für eine verlässliche Bestimmung der Reliabilität und der Validität eines Tests. Für die Durchführung der Experimente wurde deshalb eine Webapplikation entwickelt.

6 Ziele der Experimente

Bei dem Antwortbewertungsverfahren Multiple Evaluation haben die Teilnehmer im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren die Möglichkeit, ihr Wissen in Form von Antwortsicherheit differenziert wiederzugeben. Bei Unsicherheit müssen sie sich nicht für eine Antwortoption entscheiden, sondern können auch Nicht- oder Teilwissen einräumen. Die Auswertung des Tests mit einer logarithmischen Auswertefunktion, die Strafzahlungen vorsieht, stellt zudem sicher,

dass ein Teilnehmer seine Punktsumme nur dann maximieren kann, wenn er seine tatsächliche subjektive Antwortsicherheit vollkommen unverfälscht berichtet (Shuford, Albert & Massengill, 1966). Angesichts dieser Vorteile der Multiplen Evaluation war anzunehmen, dass die Varianz der beobachtbaren Testwerte eines Tests durch den Einsatz dieses Antwortverfahrens im Vergleich zum herkömmlichen Multiple-Choice-Verfahren reduziert werden kann. Die vorliegende Arbeit hatte das Ziel, die Frage empirisch zu beantworten, ob Multiple Evaluation daher ein geeigneteres Antwortverfahren zur Messung von Wissen ist, als das herkömmliche Multiple-Choice-Format. Dazu wurde mithilfe von drei Experimenten geprüft, ob die Reliabilität und die Validität eines Wissenstests durch den Einsatz des Antwortverfahrens Multiple Evaluation im Vergleich zu einem herkömmlichen Multiple-Choice-Test verbessert werden können.

Wie bereits ausgeführt, war es denkbar, dass die Teilnehmer durch eine logarithmische Auswertung nur dann zu einer unverfälschten Reproduktion ihres Wissens inzentiviert werden können, wenn sie bei der Itemvorgabe über die Punktauszahlung informiert werden. Vor diesem Hintergrund war anzunehmen, dass die Varianz der beobachtbaren Testwerte eines Tests reduziert werden kann, wenn den Teilnehmern eine solche Information zur Verfügung steht, im Vergleich zu den Bedingungen, in denen die Teilnehmer nicht über eine Auszahlungsinformation verfügen. Um diese Annahme zu überprüfen, wurde durch das erste Experiment untersucht, ob eine Verbesserung der Reliabilität eines Tests dadurch erzielt werden kann, dass die Teilnehmer über die Auszahlung informiert werden.

Beim Einsatz der Multiplen Evaluation stellte sich auch die Frage, wie differenziert Antwortsicherheit erfasst werden muss. Dabei war der Aspekt zu berücksichtigen, dass in Abhängigkeit von der Höhe der Genauigkeit eine Verschlechterung der Ökonomie des Verfahrens zu erwarten ist. Der Einsatz von Schiebereglern ermöglicht zwar eine prozentgenaue Erfassung von Antwortsicherheit, jedoch ist das Einstellen der Schieberegler für einen Teilnehmer aufwendig und daher zeitintensiv. Außerdem war davon auszugehen, dass er gar nicht in der Lage ist, seine Antwortsicherheiten prozentgenau zu differenzieren (Leclerq, 1983). Ein Antwortdreieck bietet einem Teilnehmer demgegenüber nur wenige mögliche Verteilungen der Antwortsicherheit. Die Bedienung ist jedoch im Vergleich zu den Schiebereglern einfacher, da ein Dreieck die simultane Angabe von

Antwortsicherheit in drei mögliche Antwortoptionen durch die Auswahl eines einzigen Feldes erlaubt. Deshalb lag es nahe, dass ein Dreieck eine wesentlich ökonomischere Möglichkeit ist, um Antwortsicherheit zu erheben als Schieberegler. Offen war aber, ob mit 16 Antwortmöglichkeiten Antwortsicherheit ausreichend genau erhoben werden kann, oder ob eine wesentlich differenziertere Erhebung erforderlich ist. Das erste Experiment hatte zum Ziel zu überprüfen, ob das ökonomischere Dreieck eine ausreichende Genauigkeit der Erhebung bietet, damit etwaige Vorteile des Antwortverfahrens Multiple Evaluation im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren messbar werden. Dazu wurde die Antwortsicherheit einmal mit Schiebereglern und einmal mit einem Antwortdreieck erhoben und der Einfluss der Art der Erhebung auf die Reliabilität des Tests untersucht.

Bei einer logarithmischen Auswertung (Dirkzwager, 2003) legt der Testleiter die Höhe der Strafzahlungen fest, indem er einen Wert für den Toleranzfaktor T wählt. Je höher dieser Wert ist, desto höher ist die Anzahl der Items, die ein Teilnehmer vollkommen richtig beantworten muss, um den Punkteverlust durch ein völlig falsch beantwortetes Item wieder aufzuwiegen. Beträgt der Toleranzfaktor beispielsweise T=3, so sind dazu drei Items erforderlich. Da Teilnehmer in der Regel das Ziel haben, ihre Gesamtpunktsumme in einem Test zu maximieren, war anzunehmen, dass sie dieses Risiko nicht eingehen und ihre Antwortsicherheiten umso unverfälschter berichten, je höher die zu erwartenden Strafzahlungen sind (Dirkzwager, 2003; Shuford, Albert & Massengill, 1966; Shuford & Brown, 1975). Daher sollte die Varianz der beobachtbaren Testwerte aufgrund von Teilnehmern, die ihr Wissen verfälscht wiedergeben, umso mehr verringert sein, je höher die Strafzahlungen sind. Mithilfe des zweiten Experiments wurde geprüft, ob als eine Folge daraus die Reliabilität und die Validität eines Tests mit Multipler Evaluation in Abhängigkeit von der Höhe der Strafzahlungen verbessert werden können. Bei einer linearen Auswertung wurde hingegen die geringste Testgüte erwartet, da es bei dieser Auswertung die beste Strategie eines Teilnehmers ist, immer eine absolute Antwortsicherheit in die plausibelste Antwortoption anzugeben.

Je geringer die Anzahl der Antwortoptionen eines Multiple-Choice-Tests ist, desto höher ist die Chance, die richtige Antwort zu erraten. Vor diesem Hintergrund war zu erwarten, dass die Varianz der beobachtbaren Testwerte aufgrund von

Rateprozessen bei Items mit zwei Antwortoptionen höher ist als bei Items mit drei Antwortoptionen. Daher sollten sich die Reliabilität und damit auch die Validität eines Tests mit Multiple-Choice verschlechtern, wenn die Anzahl der Antwortoptionen der Items verringert wird. Bei einem Test mit Multipler Evaluation hingegen haben die Teilnehmer die Möglichkeit, ihr Wissen differenziert zu reproduzieren, da sie auch Nicht- oder Teilwissen angeben können. Besonders wenn hohe Strafpunkte ausgezahlt werden, ist dies für einen Teilnehmer sogar die einzig sinnvolle Strategie, wenn er sein Ergebnis maximieren will. Daher sollte die Varianz der beobachtbaren Testwerte eines Tests mit Multipler Evaluation nicht durch die Anzahl der Antwortoptionen beeinflusst werden. Im dritten Experiment wurde dieser Sachverhalt empirisch überprüft und der Einfluss der Anzahl der Antwortoptionen auf die Reliabilität und die Validität eines Tests mit Multipler Evaluation im Vergleich zu einem Test mit Multiple-Choice beobachtet.

Ausschlaggebend für eine hohe psychometrische Qualität eines Tests mit Multipler Evaluation scheint die perfekte Kalibrierung der Teilnehmer zu sein. Für den Fall, dass ein Testteilnehmer nicht perfekt kalibriert ist, wurde vorgeschlagen, eine Korrektur der Antwortsicherheiten auf der Basis des individuellen Realismusindex eines Teilnehmers nachträglich vorzunehmen (Brown & Shuford, 1973; Holmes, 2002). Auf diese Weise soll die Varianz der beobachtbaren Testwerte, die durch eine verfälschte Reproduktion der Antwortsicherheiten entsteht, reduziert werden können. Die vorliegende Arbeit hatte das Ziel, zu klären, ob diese nachträgliche Korrektur deshalb ein geeignetes Verfahren ist, mit dessen Hilfe die Reliabilität und die Validität eines Tests verbessert werden können.

7 Experimentieren im World Wide Web

Alle Experimente der vorliegenden Arbeit wurden als Webexperimente durchgeführt. Reips (2003) zufolge stellen Webexperimente eine konsequente Erweiterung von Labor- und Feldexperimenten im Internet dar. Auch Buchanan und Smith (1999) waren der Ansicht, dass das Internet ein großes Potential für die psychologische Forschung biete. Die Nutzung des Internets hat den Vorteil, dass ein Testteilnehmer das Experiment jederzeit aufrufen und zuhause an seinem Computer durchführen kann. Dies erspart ihm die Anfahrt zu einem Laborraum sowie dem Testleiter die Bereitstellung des entsprechenden Raumes (Reips, 2000a). Die Region,

in der Teilnehmer rekrutiert werden können, ist räumlich nicht begrenzt, so dass Teilnehmer auf der ganzen Welt angesprochen werden können. Für den Testleiter bietet die Nutzung des Internets so den Vorteil, dass er eine große Anzahl von Teilnehmern messen kann. So erhalten seine Ergebnisse eine hohe statistische Güte, ohne enormen zeitlichen, technischen, logistischen und finanziellen Aufwand (Bandilla, 2002; Bosnjak, 2002; Reips, 2005). Während eines Webexperiments nimmt der Versuchsleiter keinen direkten Kontakt mit den Versuchspersonen auf. Dadurch wird der Versuchsleitereffekt minimiert und die Objektivität verbessert (Reips, 2005). Ein weiterer Vorteil kann die Anonymität der Datenerhebung sein. Locke und Gilbert (1995) zeigten, dass die Durchführung von Tests auf einem Computer ohne persönliche Kontakte zu einem Versuchsleiter die Bereitschaft zur Selbstauskunft erhöht. Hat eine Versuchsperson jedoch Verständnisfragen, so können diese nicht sofort durch den Versuchsleiter geklärt werden. Das Design eines Webexperiments muss deshalb sehr sorgfältig gestaltet werden. Die Aufgabe sollte für die Versuchsperson ohne weitere Erklärungen klar verständlich sein (Reips, 2000b).

Bei den ersten Webexperimenten wurde kritisiert, dass die Population der Internetnutzer bezüglich ihrer soziodemographischen Daten nicht mit der Allgemeinbevölkerung übereinstimme. Die Anzahl junger, technikinteressierter Männer überwiege. Mit der heutigen weiten Verbreitung des Internets muss eine solche Verzerrung nicht mehr befürchtet werden (Gosling, Vazire, Srivastava & John, 2004). Es ist davon auszugehen, dass eine durch ein Webexperiment erhobene Stichprobe in Merkmalen wie z.B. Alter, Geschlecht, Kultur, Bildung und sozioökonomischen Faktoren wesentlich diverser ist als eine auf herkömmliche Art selektierte Stichprobe (Buchanan & Smith, 1999). Smart (1966) recherchierte Artikel psychologischer Zeitschriften und fasste zusammen, dass die Stichproben der publizierten Experimente hauptsächlich aus männlichen Studenten der Psychologie in einer frühen Studienphase bestanden. Dies war beispielsweise in 86% der Experimente der Fall, die im „Journal of Experimental Psychology“ veröffentlicht wurden, und in 73% der Artikel des „Journal of Abnormal and Social Psychology“. Smart (1966) war der Ansicht, dass diese Universitätsstichproben sich deutlich von anderen unterschieden, beispielsweise in Alter, sozialem Status und Lernfähigkeit. Er forderte deshalb eine breitere Selektionsbasis für Versuchspersonen. Buchanan und Smith (1999) fanden 33 Jahre später in den Artikeln des „British Journal of Social

Psychology“ immer noch, dass 33 von 39 Studien Studenten als Versuchspersonen rekrutiert hatten. Die Nutzung des Internets als experimentelle Plattform bietet eine ökonomische Möglichkeit, diverse Stichproben zu messen. Die Erhebung soziodemographischer Daten kann zudem sicherstellen, dass die Stichprobe angemessen ist für die untersuchte Fragestellung (Schmidt, 1997).

Bei einem Webexperiment findet eine Selbstselektion der Versuchspersonen statt. Sie entscheiden freiwillig, am Experiment teilzunehmen, und sind am Thema des Experiments interessiert (Reips, 2000b). In der Regel verfügen die Versuchspersonen deshalb über eine hohe Motivation, das Experiment durchzuführen (Reips, 2000a). Stichproben mit speziellen Merkmalen können ausfindig gemacht und kontaktiert werden (Buchanan & Smith, 1999). Bei Studenten, die als Versuchspersonen rekrutiert werden, spielen häufig andere Gründe, wie z.B. die Pflicht, Versuchspersonenstunden abzuleisten oder eine finanzielle Entschädigung, eine Rolle für ihre Teilnahme (Buchanan & Smith, 1999). Oakes (1972) untersuchte eine Stichprobe, die aus Studenten der Psychologie bestand, im Vergleich zu einer Stichprobe, die sich aus Teilnehmern zusammensetzte, die sich aufgrund von Zeitungsanzeigen freiwillig für das Experiment gemeldet hatten. Die Selbstselektion führte zu signifikant unterschiedlichen Ergebnissen. Oakes (1972) folgerte aus seiner Studie, dass jede Art der Selektion einer Stichprobe in Abhängigkeit von der untersuchten Fragestellung zu einer begrenzten externen Validität und Generalisierbarkeit führen könne.

Den Vorwurf, die Ergebnisse von Webexperimenten seien verschieden von den mit herkömmlichen Methoden gemessenen Daten, konnten Buchanan und Smith (1999) entkräften. Sie erhoben Persönlichkeitseigenschaften mithilfe eines Webexperiments im Vergleich zu einem herkömmlichen Papier-und-Bleistift-Fragebogen. Buchanan und Smith (1999) fanden eine gute Übereinstimmung der Ergebnisse, wobei die durch das Webexperiment erhobenen Daten eine bessere Reliabilität zeigten. Diese Verbesserung erklärten sie durch eine größere Ehrlichkeit, Offenheit und Heterogenität der Teilnehmer des Webexperiments. Krantz und Dalal (2000) bestätigten ebenfalls eine gute Übereinstimmung der Ergebnisse von Webexperimenten mit Laborexperimenten.

Ein Webexperiment fordert von einem Versuchsleiter eine besonders sorgfältige Kontrolle der Versuchsbedingungen. Unbekannte konfundierende Variablen

könnten die Varianz der beobachtbaren Testwerte vergrößern und so die Datenqualität mindern (Buchanan & Smith, 1999). Da die Personen freiwillig am Experiment teilnehmen und auch keiner weiteren Kontrolle durch den Versuchsleiter unterliegen, haben sie jederzeit die Möglichkeit, das Experiment vorzeitig abzubrechen. Um diesen *Drop-out* zu reduzieren, können Techniken eingesetzt werden wie eine bedingungsunabhängige Aufwärmphase, eine Belohnung für die Teilnahme oder die Verwendung interessanten Testmaterials (Reips, 2000b). Die Motivation eines Teilnehmers, an einem Experiment teilzunehmen, kann erhöht werden z.B. durch die Gabe von direktem individuellem *Ergebnisfeedback*. Wenn die Teilnehmer wissen, dass das *Feedback* individuell nur auf der Basis ihrer Antworten erstellt wird, werden sie wahrscheinlich motivierter sein, akkurate und überlegte Antworten zu geben, was wiederum die Datenqualität erhöht (Schmidt, 1997). Verlassen in einer Versuchsbedingung besonders viele Teilnehmer vorzeitig das Experiment, so kann dies auf eine Konfundierung hinweisen (Reips, 2000b). Wird ein Experiment im Labor durchgeführt, so führen die Versuchspersonen das Experiment eher vollständig durch, besonders dann, wenn Sie durch äußere Faktoren wie die Ableistung von vorgeschriebenen Versuchspersonenstunden oder einer Bezahlung motiviert sind. Dies kann eine Kontamination der Ergebnisse von Laborexperimenten zur Folge haben (Reips, 2000a). Bei Webexperimenten können Messfehler auftreten, wenn Fehler im Design der Website bestehen (Bosnjak, 2002). Diese können jedoch durch eine sorgfältige Kontrolle vermieden werden. Die technische Varianz, die dadurch entsteht, dass die Versuchspersonen verschiedene Computer, Monitore, Webbrowser usw. nutzen, trägt zwar zur Varianz der beobachtbaren Testwerte bei, sie verhindert aber möglicherweise technisch bedingte, unentdeckte systematische Fehler, die auftreten können, wenn das Experiment nur auf einem Laborcomputer durchgeführt wird. Diese Varianz bietet zudem den Vorteil, dass die Ergebnisse besser generalisiert werden können (Reips, 2000b). Probleme wie unvollständige oder unangemessene Antworten sowie Mehrfachteilnahmen können durch ein gut durchdachtes Design der Webapplikation weitgehend verhindert werden (Schmidt, 1997). Außerdem sollte eine Website für potentielle Teilnehmer durch ein ansprechendes Design und gute Funktionalität attraktiv sein (Reips, 2001).

8 Testmaterial

Als Testmaterial wurde ein Englischtest in Anlehnung an den Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) eingesetzt. Dieser Referenzrahmen wird von Harsch (2005, S. 3) wie folgt dargestellt:

> *Mit seinem Referenzsystem stellt der GER ein Kompetenzmodell bereit, das relevante Teilbereiche kommunikativen Handelns und sprachlichen Könnens kategorisiert und beschreibt.*

Der Gemeinsame Europäische Referenzrahmen für Sprachen teilt Sprachfähigkeiten in die folgenden sechs Stufen ein: A1, A2, B1, B2, C1, C2. Dabei sind Items der Stufen A1 und A2 mit geringen Kenntnissen lösbar, B1 und B2 mit fortgeschrittenen Fähigkeiten und die Stufen C1 und C2 nur mit sehr guten Sprachkenntnissen. Der in den Experimenten verwendete Test bestand insgesamt aus 36 Items mit je einer richtigen Antwort und zwei Distraktoren. Für jede Schwierigkeitsstufe des GER wurden sechs Items konstruiert. Der Test beinhaltete also zwölf Items, deren Schwierigkeit der GER-Stufe A entsprachen, zwölf Items, die die Schwierigkeit der Stufe B, und zwölf Items, die die Schwierigkeit der Stufe C hatten. Die Items der Stufe A werden im Folgenden als einfach, die der Stufe B als mittelschwierig und die der Stufe C als schwierig bezeichnet. In einem Vorexperiment wurden die Items mithilfe von ca. 7000 Teilnehmern normiert. Für den in den Experimenten verwendeten Englischtest wurden nur die Items ausgewählt, die mindestens eine *part-whole-korrigierte* Trennschärfe von 0,20 zeigten und deren Lösungswahrscheinlichkeit der jeweiligen Schwierigkeitsstufe des Gemeinsamen Europäischen Referenzrahmens für Sprachen entsprach. Der Englischtest wurde den Teilnehmern als kostenloser Test der Heinrich-Heine-Universität Düsseldorf im Internet zur Verfügung gestellt. Als Anreiz für ihre Teilnahme erhielten die Teilnehmer im Anschluss an den Test eine detaillierte Einstufung ihrer Fähigkeiten auf der Basis der Stufen des Gemeinsamen Europäischen Referenzrahmens.

9 Erstes Experiment

9.1 Fragestellungen und Hypothesen des ersten Experiments

Im Gegensatz zu einem Test mit Multiple-Choice kann ein Teilnehmer bei einem Test mit Multipler Evaluation sein Wissen differenziert in Form seiner tatsächlichen subjektiven Antwortsicherheit berichten. Bei der Multiplen Evaluation kann er damit Teilwissen angeben und auch Nichtwissen einräumen. Die Auswertung mit einer logarithmischen Funktion soll einen Teilnehmer zudem durch Strafzahlungen inzentivieren, seine tatsächlichen subjektiven Antwortsicherheiten vollkommen unverfälscht zu berichten. Das erste Experiment untersucht die Frage, ob die Multiple Evaluation deshalb besser geeignet ist das Wissen eines Teilnehmers zu quantifizieren als Multiple-Choice. Dazu wurde untersucht, ob durch den Einsatz des Antwortverfahrens Multiple Evaluation die Reliabilität eines Tests im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren verbessert werden kann. Dies wurde anhand der folgenden Hypothese geprüft.

I. Hypothese:
Wenn die Messung von Englischfähigkeiten mit Multipler Evaluation und einer logarithmischen Auswertung erfolgt, verbessert sich die interne Konsistenz des Tests im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren**.**

Beim Einsatz des Antwortverfahrens Multiple Evaluation stellte sich die Frage, wie differenziert Antwortsicherheit erhoben werden muss, damit etwaige Vorteile des Verfahrens im Vergleich zu Multiple-Choice messbar werden. Nach De Finetti (1965) sollte eine kontinuierliche Skala verwendet werden, denn diese könne eine größere Menge an Informationen messen als eine diskrete. Leclerq (1983) war dagegen der Ansicht, dass eine zu große Genauigkeit nicht sinnvoll sei. Für die Testteilnehmer sei es zu schwierig, ihre Antwortsicherheit prozentgenau anzugeben. Empirische Befunde deuten darauf hin, dass schon nur wenige diskrete Möglichkeiten, Antwortsicherheit auf die Antwortoptionen zu verteilen, ausreichend sind, um die psychometrische Qualität eines Tests zu erhöhen. Bokhorst (1986) zeigte

beispielsweise, dass die Reliabilität eines Multiple-Choice-Tests schon durch den Einsatz einer biopolaren verbalen Skala verbessert werden kann. Abedi und Bruno (1989) beobachteten eine deutliche Verbesserung der Reliabilität eines Tests mit Multipler Evaluation beim Einsatz eines Dreiecks, das 13 mögliche Verteilungen der Antwortsicherheit repräsentierte. Angesichts dieser Befunde war anzunehmen, dass eine Erhebung von Antwortsicherheit mit nur wenigen diskreten Antwortmöglichkeiten ausreichend genau ist, um die Güte eines Tests im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren zu verbessern. Im ersten Experiment wurde geprüft, ob eine prozentgenaue Erhebung von Antwortsicherheit dennoch zu einer höheren Reliabilität eines Tests führt als eine Erhebung mit nur wenigen diskreten Antwortmöglichkeiten. Dazu wurde die folgende Hypothese formuliert:

II. Hypothese:
Wenn die Erhebung von Antwortsicherheit in einem Englischtest mit Multipler Evaluation durch Schieberegler erfolgt, verbessert sich die interne Konsistenz gegenüber der Erhebung von Antwortsicherheit mit einem Antwortdreieck.

Ein Antwortinstrument mit nur 16 Antwortmöglichkeiten wurde dabei durch ein gleichseitiges Dreieck, das in 16 Antwortfelder unterteilt war, realisiert (Paul, 1993). Eine prozentgenaue Erfassung von Antwortsicherheit erfolgte mithilfe von Schiebereglern. Jeder Antwortoption war dabei ein Schieberegler zugeordnet. Diese Schieberegler konnten auf ganzzahlige Werte zwischen 0% und 100% in Schritten von 1% eingestellt werden. Bei drei Antwortoptionen waren so 5151 verschiedene Verteilungen von Antwortsicherheit möglich. Die Berechnung der Anzahl der möglichen Verteilungen wird in Anhang F dargestellt.

Bei Tests mit Multipler Evaluation wurde häufig die verlängerte Durchführungszeit gegenüber Multiple-Choice-Tests kritisiert (Ebel, 1968; Koehler, 1971; Rippey, 1968b). Im ersten Experiment wurde die Durchführungszeit gemessen, um zu untersuchen, ob dieser Nachteil trotz der beträchtlichen Entwicklung der Computertechnik auch heute noch besteht. Ein Vorteil des Antwortdreiecks nach Paul (1993) ist dessen einfache Bedienbarkeit. Um ein Item zu beantworten, ist nur ein Klick mit der Maustaste in ein Feld erforderlich. Bei der Beantwortung eines Items mit den Schiebereglern mussten die Teilnehmer hingegen in der Regel

mindestens einen Schieberegler bewegen und danach eine Schaltfläche betätigen, um ihre Antwort abzugeben. Deshalb wurde eine wesentlich kürzere Durchführungszeit mit dem Dreieck erwartet, im Vergleich zur Testdurchführung mit den Schiebereglern.

Eine weitere Fragestellung des Experiments war, ob Teilnehmer durch eine logarithmische Auswertung, die Strafzahlungen vorsieht, nur dann zu einer unverfälschten Reproduktion ihrer Antwortsicherheiten inzentiviert werden können, wenn sie bei der Bearbeitung eines Items über die Punktauszahlung informiert werden. Es war also zu prüfen, ob durch eine Information über die Auszahlung die Varianz der beobachtbaren Testwerte, erzeugt durch eine verfälschte Wiedergabe der tatsächlichen subjektiven Antwortsicherheiten, reduziert werden kann. Deshalb wurde untersucht, ob die Reliabilität eines Tests durch eine Information über die Auszahlung verbessert werden kann, im Vergleich zu den Testbedingungen, in denen die Teilnehmer nicht über diese Information verfügen. Dazu wurde die folgende Hypothese formuliert:

III. Hypothese:
Wenn die Teilnehmer des Englischtests mit Multipler Evaluation eine Information über die Auszahlung erhalten, dann verbessert sich die interne Konsistenz des Tests im Vergleich zu den Bedingungen mit Multipler Evaluation, in denen diese Information nicht zur Verfügung steht.

In der Kontrollbedingung mit Multiple-Choice wurde hingegen keine Verbesserung der Reliabilität durch die Auszahlungsinformation erwartet. Die Testteilnehmer hatten auch mit dieser Information keine andere Handlungsoption als zu raten, wenn sie unsicher waren. Zudem wurde Raten nicht durch negative Punktauszahlungen bestraft. Raten war also die beste Teststrategie für einen Teilnehmer, um seine Punktsumme zu maximieren. Daher wurde die folgende Hypothese abgeleitet:

IV. Hypothese:
Wenn die Teilnehmer des Englischtests mit Multiple-Choice eine Information über die Auszahlung erhalten, so zeigt sich keine signifikante Verbesserung der internen

Konsistenz des Tests im Vergleich zu der Bedingung mit Multiple-Choice, in denen diese Information nicht zur Verfügung steht.

Wie bereits ausgeführt, ist bei der Durchführung eines Tests davon auszugehen, dass nicht alle Teilnehmer perfekt kalibriert sind. Durch eine Korrektur der berichteten Antwortsicherheiten mithilfe eines individuellen Realismusindex soll die Varianz der beobachtbaren Testwerte durch Teilnehmer, die ihre Antwortsicherheit aufgrund von *Over-* oder *Underconfidence* verfälscht wiedergeben, nachträglich reduziert werden können. Im ersten Experiment wurde mithilfe der folgenden V. Hypothese geprüft, ob diese Korrektur deshalb ein geeignetes Verfahren ist, um die Reliabilität eines Tests zu verbessern.

V. Hypothese:
Werden die Antwortsicherheiten der Testteilnehmer auf der Basis ihres individuellen Realismusindex korrigiert, so verbessert sich die interne Konsistenz, ermittelt auf der Basis der erneuten Punktauszahlung, im Vergleich zur internen Konsistenz, ermittelt für die Punkte, die anhand der nicht korrigierten Antwortsicherheiten ausgezahlt wurden.

9.2 Methode

Die Experimente dieser Arbeit wurden als Webexperimente durchgeführt. Dem Design des Versuchsplans und den Anforderungen des Experiments entsprechend wurde eine Webapplikation mit Datenbankanbindung erstellt.

9.2.1 Design

Im ersten Experiment wurde die erste unabhängige Variable „Antwortinstrument" für den Test mit Multipler Evaluation in den Stufen „Antwortdreieck" und „Schieberegler" untersucht. Das Antwortdreieck stellte einem Teilnehmer 16 diskrete Antwortmöglichkeiten zur Verfügung. Die Schieberegler ermöglichten eine auf 1%

Tabelle 9-1

Versuchsdesign des ersten Experiments.

	Auszahlungsinformation	
Antwortinstrument	ohne	mit
Multiple Evaluation Dreieck		
Multiple Evaluation Schieberegler		
Multiple-Choice		

genaue Angabe von Antwortsicherheit. Die Bedingung mit Multiple-Choice wurde durch Optionsfelder (*Radiobuttons*) realisiert. Die zweite unabhängige Variable „Auszahlungsinformation" war eine Information des Teilnehmers über die Auszahlung in den Stufen „Testdurchführung mit Auszahlungsinformation" und „Testdurchführung ohne Auszahlungsinformation". Das Versuchsdesign zeigt Tabelle 9-1.

Das Testmaterial war ein Englischtest, der aus je zwölf einfachen (GER-Stufe A), mittelschwierigen (GER-Stufe B) und schwierigen Items (GER-Stufe C) bestand (siehe Abschnitt 8). Da die Dreiecksfigur auf genau drei Antwortoptionen limitiert ist, wurden Items mit drei Antwortoptionen, d.h. einer richtigen Antwort und zwei Distraktoren, verwendet. Um den Einfluss der Schwierigkeit zu untersuchen, wurden die Ergebnisse zusätzlich zu der Betrachtung der Items des

gesamten Tests auch getrennt für einfache, mittelschwierige und schwierige Items ermittelt.

9.2.2 Auswertung des Englischtests

Die Auszahlung der Punkte für die Bedingungen mit Multipler Evaluation erfolgte anhand einer logarithmischen Funktion nach Dirkzwager (2003) mit dem Toleranzfaktor T=1 (t=0,1667). Damit betrug der Auszahlungbereich -100 bis 100 Punkte. Die Auszahlung der Punkte erfolgte anhand der folgenden Funktion:

$$S(r_c(i)) = \frac{\ln[(1-0{,}1667*3)*r_c(i)+0{,}1667] + \ln(3)}{\ln(1-0{,}1667*3+0{,}1667) + \ln(3)}$$

Dabei bedeutet:
S= Punktauszahlung
r_c= Antwortsicherheit in die richtige Antwortoption
i= Item, i=1, …, 36

Tabelle 9-2 zeigt die Punktauszahlungen für alle Antwortmöglichkeiten des Dreiecks und für einige prozentuale Antwortsicherheiten, die mit den Schiebereglern eingestellt werden konnten. In den Bedingungen mit Multiple-Choice erfolgte die Auswertung dichotom. Die Teilnehmer erhielten einen Punkt für eine richtige und null Punkte für eine falsche Antwort.

Tabelle 9-2

Punktauszahlung für alle Antwortmöglichkeiten des Dreiecks und ausgewählte Prozentwerte der Schieberegler.

Prozent Antwortsicherheit in die richtige Antwort	0%	10%	15%	20%	33,3%	40%	50%	60%	70%	80%	90%	100%
Punktauszahlung Dreieck	-100	-	-46	-32	0	-	32	-	63	77	-	100
Punktauszahlung Schieberegler	-100	-62	-46	-32	0	14	32	49	63	77	89	100

9.2.3 Operationalisierung der unabhängigen Variablen „Antwortinstrument“

Das Antwortdreieck war ein gleichseitiges, auf einer Seite stehendes Dreieck. Die Ecken dieses Dreiecks waren im Gegenuhrzeigersinn mit den Buchstaben A, B und C gekennzeichnet. Jeder Ecke war eine Antwortoption des Items zugeordnet. Diese wurde direkt neben der jeweiligen Ecke angezeigt. Die Antwortoptionen wurden in einer der Farben Rot, Grün oder Blau dargestellt. Die Fläche des Dreiecks war in 16 Felder unterteilt, und jedes Feld entsprach einer bestimmten Verteilung von Antwortsicherheit auf die drei Antwortoptionen. Dabei ergab die Summe der Antwortsicherheiten immer 100%. Welche Verteilung von Antwortsicherheiten ein Feld repräsentierte, wurde durch seine Position bestimmt und durch seine Farbe verdeutlicht. Zur Einfärbung der Felder wurden Farben der additiven Farbmischung verwendet. Das dunkelste Feld der jeweiligen Farbe, an einer Ecke des Dreiecks, bedeutete eine Angabe von 100% Antwortsicherheit in diese Antwortoption und von jeweils 0% in die beiden anderen Optionen. Je weiter ein Feld im Dreieck von einer Ecke entfernt war, desto geringer war die Antwortsicherheit, mit der ein Testteilnehmer sich für diese Antwortoption entschied. Die Felder, die genau in der Mitte zwischen zwei Antwortoptionen lagen, wurden in den jeweiligen Mischfarben der additiven Farbmischung dargestellt und repräsentierten eine Antwortsicherheit

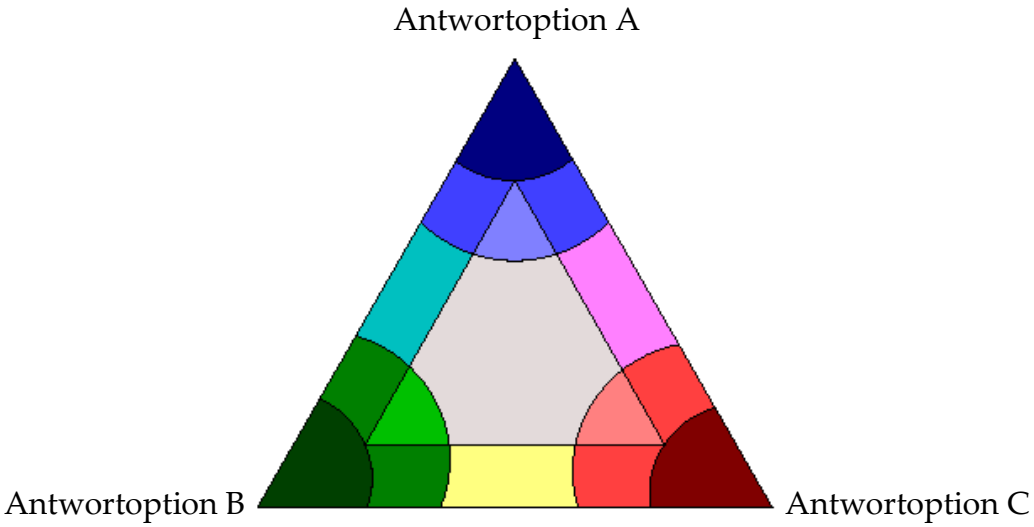

Abbildung 9-1: Antwortdreieck, in dem alle Felder farbig dargestellt sind.

Tabelle 9-3

Verteilungen der Antwortsicherheit auf drei Antwortoptionen und die Farben der 16 Felder des Dreiecks.

Nr.	Antwort A	Antwort B	Antwort C	Farbe
1	100%	0%	0%	Dunkelblau
2	80%	20%	0%	Mittelblau
3	80%	0%	20%	Mittelblau
4	70%	15%	15%	Hellblau
5	50%	50%	0%	Türkis
6	50%	0%	50%	Magenta
7	33,3%	33,3%	33,3%	Grau
8	20%	80%	0%	Mittelgrün
9	20%	0%	80%	Mittelrot
10	15%	70%	15%	Hellgrün
11	15%	15%	70%	Hellrot
12	0%	100%	0%	Dunkelgrün
13	0%	80%	20%	Mittelgrün
14	0%	50%	50%	Gelb
15	0%	20%	80%	Mittelrot
16	0%	0%	100%	Dunkelrot

von 50% in jede der beiden Antwortoptionen. Das mittlere Feld bedeutete eine Angabe von Antwortsicherheit von gerundet je 33% in alle drei Antwortoptionen. Dieses Feld wählte ein Testteilnehmer, wenn er völliges Nichtwissen ausdrücken wollte. Abbildung 9-1 zeigt ein Antwortdreieck in dem alle Felder farbig dargestellt sind. Tabelle 9-3 beinhaltet alle 16 möglichen Verteilungen von Antwortsicherheit auf drei Antwortoptionen und die jeweiligen Farben der Felder. In der Testapplikation wurden bei der Vorgabe eines Items im Antwortdreieck nur die Rahmen der Felder angezeigt. Erst wenn ein Teilnehmer mit dem Mauszeiger über das Dreieck fuhr, wurde das jeweils überfahrene Feld farbig. Abbildung 9-2 zeigt das Dreieck aus der Testapplikation, nachdem ein Feld ausgewählt wurde. Dem Teilnehmer wurde so verdeutlicht, welches Feld des Dreiecks er aktuell ausgewählt hatte. Gleichzeitig wurde die Anzeige der prozentualen Antwortsicherheit an den Ecken des Dreiecks aktualisiert. Klickte ein Testteilnehmer mit der linken Maustaste in ein Feld des Dreiecks, so wurde die dem Feld entsprechende Verteilung der

Antwortsicherheit auf die drei Antwortoptionen gespeichert und der Teilnehmer zur nächsten Seite weitergeleitet.

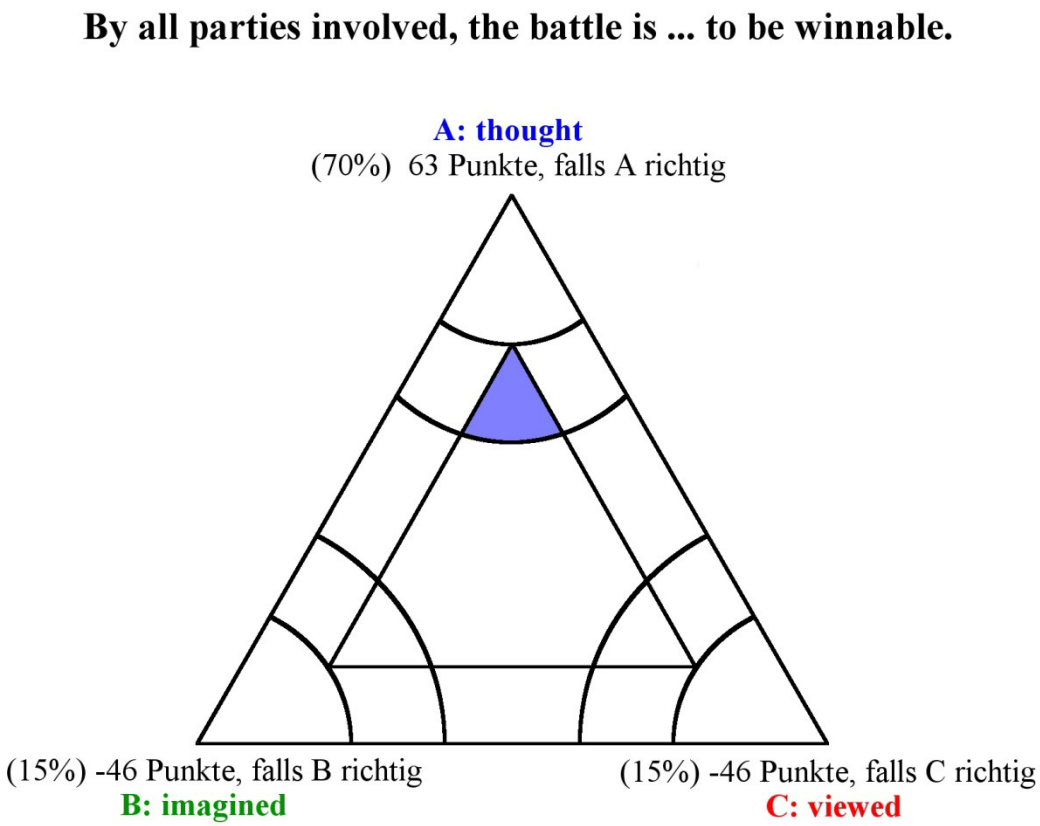

Abbildung 9-2: Der Teilnehmer hat 70% Antwortsicherheit in die Option A angegeben und jeweils 15% in die Optionen B und C.

Technisch wurde das Antwortdreieck durch ein *GIF*-Bild umgesetzt, welches mithilfe einer *HTML-Image-Map* in 16 einzelne Segmente unterteilt war. Diese Segmente waren jeweils mit den *JavaScript-Events onmouseover*, *onmouseout* und *onclick* verknüpft, die je nach Interaktion des Testteilnehmers verschiedene Funktionen ausführten.

Für das aus Schiebereglern bestehende Instrument wurden drei horizontale Regler übereinander angeordnet und mit den Buchstaben A, B und C gekennzeichnet. Links neben einem Schieberegler wurde die zugeordnete Antwortoption angezeigt. In der Ausgangsposition waren alle Schieberegler auf den Wert 33% eingestellt. Die Regler waren so miteinander verbunden, dass die drei

Einzelwerte in der Summe immer 100% ergaben. Abbildung 9-3 zeigt die Schieberegler mit den Einstellungen: Schieberegler A=70%, Regler B=15% und Regler C=15%. Ein Testteilnehmer beantwortete ein Item, indem er den Knopf eines Schiebereglers mit der Maus nach links schob, um den eingestellten Wert zu verringern, oder nach rechts, um ihn zu erhöhen. Wenn die Einstellung des zuerst bewegten Schiebereglers abgeschlossen war und ein weiterer Schieberegler bewegt wurde, wurde der zuerst bewegte Schieberegler gesperrt, d.h. der eingestellte Wert, z.B. 70%, veränderte sich nicht mehr. So konnte der Testteilnehmer nur noch die verbleibenden 30% Antwortsicherheit auf die anderen beiden Schieberegler aufteilen. Ein gesperrter Schieberegler wurde durch einen roten Hintergrund gegenüber einem nicht gesperrten, grün hinterlegten Schieberegler kenntlich gemacht. Wollte der Testteilnehmer seine Auswahl wieder aufheben, hatte er dazu zwei Möglichkeiten:

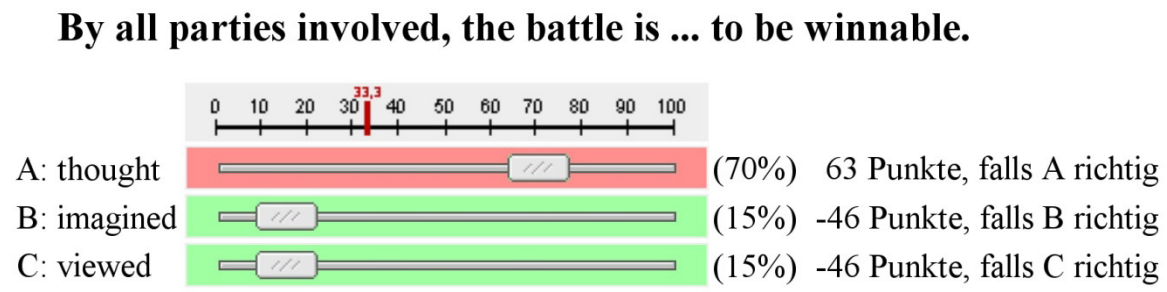

Abbildung 9-3 Schieberegler mit einer Einstellung von 70% Antwortsicherheit in die Antwortoption A und von je 15% in die Antwortoptionen B und C. Der rot hinterlegte Schieberegler ist gesperrt.

Er konnte auf eine Schaltfläche mit der Beschriftung „auf 33% zurücksetzen“ klicken, und alle drei Schieberegler wurden automatisch auf den Wert 33% zurückgesetzt. Oder er konnte Optionsfelder, die rechts neben den Schiebereglern platziert waren, deaktivieren und so die Sperrung eines Schiebereglers aufheben. Jeder einzelne Schieberegler konnte auf einen ganzzahligen Wert von 0% bis 100% eingestellt werden. Insgesamt bestanden 5151 Kombinationsmöglichkeiten, Antwortsicherheit in drei Antwortoptionen anzugeben (siehe Anhang F). Da eine ganzzahlige gleichmäßige Aufteilung von 100% auf drei Antwortoptionen bei einigen Kombinationen nicht möglich ist, ergaben sich zum Teil Abweichungen in der Größe

von zweimal 0,5%. Wurde z.B. ein Schieberegler von 100% auf 99% verschoben, so konnte das verbleibende eine Prozent nicht gleichmäßig ganzzahlig auf die beiden anderen Schieberegler aufgeteilt werden. In diesem Fall wurden die Werte 99%/0%/0% angezeigt und ggfs. gespeichert. Die aktuell eingestellten prozentualen Antwortsicherheiten wurden rechts neben den Schiebereglern angezeigt. Um eine Antwort zu speichern und zur nächsten Seite zu gelangen, musste der Teilnehmer die Schaltfläche „Antwort abgeben“ betätigen.

Technisch wurden die Schieberegler mit einer *JavaScript-Library* realisiert. Ein Schieberegler erzeugte im Programm bei jeder Interaktion des Testteilnehmers ein *JavaScript-Event*. Durch einen *JavaScript-Befehl* konnte die Position des Schiebereglers programmtechnisch auf jeden Wert zwischen 0% und 100% eingestellt werden. Die Kombination dieser *JavaScript-Eigenschaften* ermöglichte es, beim Verschieben eines der drei Schieberegler die anderen beiden so zu verändern, dass die Summe der prozentualen Antwortsicherheiten immer 100% (bzw. 99% bei Abrundung, s.o.) ergab. Bei jeder Bewegung der Schieberegler wurden die prozentualen Antwortsicherheiten aktualisiert.

Das Multiple-Choice-Antwortinstrument wurde durch drei Optionsfelder realisiert. Ein solches Optionsfeld befand sich rechts neben jeder Antwortoption. In der Ausgangssituation war keines der Felder markiert. Ein Testteilnehmer sollte das Optionsfeld neben der Antwort, die er favorisierte, mit der linken Maustaste markieren. Danach musste er seine Auswahl durch Klicken der Schaltfläche „Antwort abgeben“ bestätigen, um zur nächsten Seite zu gelangen. Diese Schaltfläche war unter den drei Antwortmöglichkeiten platziert und zunächst deaktiviert. Sie wurde erst aktiv, nachdem der Teilnehmer eines der drei

Frage:	By all parties involved, the battle is ... to be winnable.	Auswahl
A:	thought	○
B:	imagined	◉
C:	viewed	○

Abbildung 9-4: Multiple-Choice-Instrument. Antwortoption B ist ausgewählt.

Optionsfelder ausgewählt hatte. So wurde verhindert, dass ein Teilnehmer zur nächsten Seite gelangen konnte, ohne das Item bearbeitet zu haben. Abbildung 9-4 zeigt das Antwortinstrument der Bedingungen mit Multiple-Choice.

Technisch wurde das Multiple-Choice-Instrument mit den Browser-eigenen HTML-Elementen umgesetzt. Mit dem Element *<INPUT TYPE=RADIO ...>* wurden drei Optionsfelder erzeugt, von denen immer nur eines markiert werden konnte. Um die Bedienung zu erleichtern, wurde bei den Antwortoptionen die HTML-Eigenschaft *Label* verwendet. So wurde bereits bei einem Mausklick auf eine der drei Antwortoptionen das zugehörige Optionsfeld aktiviert.

9.2.4 Operationalisierung der unabhängigen Variable „Auszahlungsinformation“

Als Auszahlungsinformation erhielten die Teilnehmer in den Bedingungen mit Multipler Evaluation ein *Feedforward* und ein *Feedback*. Das *Feedforward* bestand aus der zu erwartenden Punktauszahlung. Beim Antwortdreieck erfolgte die Anzeige des *Feedforward* an den Ecken (neben den prozentualen Antwortsicherheiten) und bei den Schiebereglern rechts neben den Antwortoptionen. Die Anzeige des *Feedforward* wurde aktualisiert, sobald ein Teilnehmer die Einstellung der Schieberegler veränderte. Das *Feedback* erhielten die Teilnehmer auf einer separaten Seite, direkt nachdem sie ein Item bearbeitet hatten. In den Bedingungen mit Multipler Evaluation bestand das *Feedback* aus den folgenden Komponenten:

- Den aktuell für ein Item ausgezahlten Punkten.
- Der richtigen Antwortoption.
- Den prozentualen Antwortsicherheiten, die der Testteilnehmer den einzelnen Optionen zugeordnet hatte.
- Dem aktuellen Gesamtpunktestand.
- Zusätzlich wurde der Teilnehmer durch Kommentare nach mehrfachen negativen Punktauszahlungen darauf hingewiesen, dass er diese Strafzahlungen nur vermeiden könne, indem er seine Antwortsicherheiten unverfälscht angäbe.

In den Versuchsbedingungen mit Multipler Evaluation, in denen die Teilnehmer keine Auszahlungsinformation erhielten, wurden während der

Bearbeitung eines Items nur die aktuell eingestellten prozentualen Antwortsicherheiten angezeigt. Ein *Feedback* erhielten die Teilnehmer nicht.

In den Bedingungen mit Multiple-Choice wurden die Teilnehmer vor dem Test darüber informiert, dass ihnen bei einer richtigen Antwort ein Punkt und bei einer falschen kein Punkt ausgezahlt würde. Ein *Feedforward* erhielten sie nicht. In der Bedingung „Multiple-Choice mit Auszahlungsinformation" wurde ihnen nach jedem Item ein *Feedback* präsentiert, das aus der richtigen Antwort, der Information, ob die Auszahlung null oder einen Punkt betragen hatte sowie dem aktuellen Gesamtpunktestand bestand.

9.2.5 Operationalisierung der abhängigen Variablen

Als abhängige Variablen wurden die folgenden Werte je Teilnehmer erhoben:

- **Prozentuale Anwortsicherheit**: Zu jedem Item wurde die prozentuale Antwortsicherheit in die richtige Antwort und in die Distraktoren erfasst. Die prozentuale Antwortsicherheit in die Distraktoren wurde für die Berechnung des individuellen Realismusindex eines Testteilnehmers benötigt.
- **Punktauszahlung je Item**: Aus der prozentualen Antwortsicherheit wurde über eine *JavaScript-Funktion* im Browser die Punktauszahlung errechnet und gespeichert.
- **Bearbeitungszeiten**: Um die Zeiten ermitteln zu können, die ein Testteilnehmer für die Bearbeitung einzelner Abschnitte im Experiment benötigt hatte, wurden der Zeitpunkt des Beginns der Anzeige eines Testteils und der Zeitpunkt des Abschlusses dieses Teils gespeichert. Die jeweilige Zeitdauer wurde aus der Differenz dieser beiden Zeitpunkte errechnet.

9.2.6 Operationalisierung des individuellen Realismusindex

Der individuelle Realismusindex a eines Teilnehmers wurde nach der folgenden Formel ermittelt (Holmes, 2002, S. 132):

$$a = \frac{k^* \sum_{i=1}^{m} r_c(i) - m}{k^* \sum_{i=1}^{m} \sum_{j=1}^{k} r(i,j)^2 - m}$$

Dabei bedeutet:

a= Realismusindex

m= Anzahl der Items

k= Anzahl der Antwortoptionen

r_c= prozentuale Antwortsicherheit in die richtige Antwort (in Wahrscheinlichkeiten null bis eins)

r= prozentuale Antwortsicherheit in einen Distraktor (in Wahrscheinlichkeiten null bis eins)

i= Item, i=1, …, m

j= Antwortoption des Items, j=1, …, k

Erzielte ein Testteilnehmer einen Realismusindex von a=1, so war er im Test perfekt kalibriert. War a<1, dann zeigte der Teilnehmer *Overconfidence*, war a>1, so zeigte er *Underconfidence*.

9.2.7 Operationalisierung der Realismuskorrektur

Auf der Basis des individuellen Realismusindex a wurden die prozentuale Antwortsicherheit eines Teilnehmers in die richtige Antwort für jedes Item korrigiert. Der Teilnehmer erhielt dabei die Antwortsicherheit, die er hätte wiedergeben müssen, wenn er im Test perfekt kalibriert gewesen wäre. Diese Korrektur wurde anhand der folgenden Formel durchgeführt (Holmes, 2002, S. 132):

$$r_{korrigiert} = a*r_c+(1-a)/k$$

Dabei bedeutet:

$r_{korrigiert}$= Realismusindex-korrigierte Antwortsicherheit in die richtige Antwort (in Wahrscheinlichkeiten von null bis eins)

a= Realismusindex des Testteilnehmers

r_c= prozentuale Antwortsicherheit in die richtige Antwort (in Wahrscheinlichkeiten null bis eins)

k= Anzahl der Antwortoptionen

Auf der Basis der korrigierten Antwortsicherheiten wurden die Punkte erneut ausgezahlt.

9.2.8 Testapplikation

Zur Durchführung der Experimente wurde eine Applikation, basierend auf den Programmiersprachen *PHP* und *JavaScript* sowie der Formatierungssprache *HTML,* entwickelt. Die erhobenen Daten wurden in einer *MySQL-Datenbank* gespeichert. Die Testapplikation bestand aus den folgenden sechs Teilen:

Teil 1: Startseite

Auf der Startseite wurden die Testteilnehmer über die Rahmenbedingungen des Experiments, den voraussichtlichen, für ihre Teilnahme erforderlichen Zeitaufwand und ihre Aufgaben informiert. Der Kopfbereich der Startseite enthielt die Namen der Versuchsleiter sowie die Möglichkeit, ein Kontaktformular zu öffnen und darüber Fragen und Anmerkungen an die Verantwortlichen zu senden. Durch Betätigen der Schaltfläche „Ich möchte am Test teilnehmen“ erklärten die Testteilnehmer ihren Willen zur Teilnahme am Experiment und ihr Einverständnis für die anonyme Speicherung ihrer Daten. Nachdem die Teilnehmer auf diese Schaltfläche geklickt hatten, wurden sie zum nächsten Teil der Applikation weitergeleitet.

Teil 2: Fragebogen

Mithilfe eines Fragebogens wurden Alter, Geschlecht und Muttersprache der Testteilnehmer erfasst. Die Beantwortung der Fragen erfolgte entweder durch die Aktivierung eines Optionsfeldes (Geschlecht, Muttersprache) oder durch die Auswahl eines Wertes aus einer vorgegebenen Liste (Alter). Betätigte ein Testteilnehmer die Schaltfläche „weiter“ am Ende der Seite, so wurde zunächst geprüft, ob er alle Fragen beantwortet hatte. War dies nicht der Fall, wurden die entsprechenden Felder rot markiert und der Teilnehmer gebeten, alle Felder auszufüllen. Ein Wechsel zur nächsten Seite der Anwendung ohne die Beantwortung aller Fragen war nicht möglich. So wurde die Speicherung eines unvollständigen Datensatzes verhindert.

Teil 3: Einführung in das Antwortverfahren

Die Startseite und der Fragebogen waren für die Teilnehmer aller Bedingungen gleich. Die Einführung in das Antwortverfahren war den Antwortinstrumenten entsprechend unterschiedlich gestaltet. In den Bedingungen mit Multipler Evaluation wurde die Angabe von Antwortsicherheit mit dem jeweiligen Antwortinstrument sowie die Auswertung erklärt. Die Testteilnehmer wurden dabei darauf hingewiesen, dass sie ihre Punktsumme nur dann maximieren könnten, wenn sie ihre Antwortsicherheit vollkommen unverfälscht wiedergäben. Befand sich ein Testteilnehmer in einer Versuchsbedingung, die eine Information über die Punktauszahlung erhielt, wurden das *Feedforward* und das *Feedback* ebenfalls erläutert.

Teil 4: Übungen

Die Teilnehmer der Versuchsbedingungen mit Multipler Evaluation bearbeiteten fünf Übungsfragen, um sich mit dem Umgang mit dem jeweiligen Antwortinstrument vertraut zu machen. Die Texte der Übungen befinden sich in Anhang E. Wegen der allgemeinen Bekanntheit des Verfahrens führten die Teilnehmer der Versuchsbedingungen mit Multiple-Choice keine Übungen durch.

Teil 5: Englischtest

Bei dem Englischtest wurde jeweils ein Item zur Bearbeitung mit dem jeweiligen Antwortinstrument auf dem Bildschirm dargestellt. Im Kopfbereich dieser Seite wurde der Teilnehmer über seinen Fortschritt im Test durch die Anzeige der Nummer des Items informiert. In den Versuchsbedingungen, in denen die Teilnehmer eine Information über die Auszahlung erhielten, wurde dort auch der Gesamtpunktestand ausgegeben. Erhielt ein Teilnehmer *Feedback*, so wurde dieses auf einer separaten Seite direkt im Anschluss an die Bearbeitung jedes Items ausgegeben.

Teil 6: Testergebnis

Nach Abschluss des Englischtests erhielt jeder Testteilnehmer sein persönliches Ergebnis. Dieses bestand aus:

- Der Gesamtpunktsumme.
- Der Punktauszahlung für jedes Item als Balkendiagramm.
- In den Bedingungen mit Multipler Evaluation der durchschnittlichen prozentualen Antwortsicherheit in die richtige Antwort je Schwierigkeitsstufe des GER als Balkendiagramm.
- In den Bedingungen mit Multiple-Choice den Prozent richtig beantworteter Items je Schwierigkeitsstufe des GER als Balkendiagramm.

Unter diesem persönlichen Ergebnis wurde die Bedeutung der Schwierigkeitsstufen des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER) in Form einer Tabelle erläutert.

9.2.9 Zuordnung der Teilnehmer zu den Bedingungen

Für die randomisierte Zuordnung der Testteilnehmer zu den Versuchsbedingungen wurde folgendes Verfahren verwendet: Immer wenn ein Teilnehmer die Startseite aufrief, wurde die aktuelle Anzahl N der als auswertbar bewerteten Teilnehmer seit Beginn des Experiments pro Bedingung ermittelt. Zu dieser Anzahl wurde je Bedingung eine Zufallszahl Z zwischen null und fünf addiert. Diese Zufallszahl war eine Dezimalzahl ohne Begrenzung der Nachkommastellen. Mithilfe der Summe aus

Teilnehmerzahl und Zufallszahl wurde eine Rangreihenfolge der Versuchsbedingungen gebildet. Der neue Teilnehmer des Experiments wurde dann

Tabelle 9-4

Beispiel einer randomisierten Zuordnung von Teilnehmern zu fünf Versuchsbedingungen.

Bedingung	aktuelles N	Zufallszahl Z	Summe N+Z	Rang
A	50	0,4	50,4	2
B	48	1,9	49,9	1
C	47	4,7	51,7	3
D	51	2,1	53,1	5
E	50	1,9	51,9	4

der Bedingung mit dem geringsten Rang zugeordnet. Tabelle 9-4 zeigt ein Beispiel für fünf Versuchsbedingungen. In diesem Beispiel würde der neue Teilnehmer der Bedingung B zugeordnet. Betrug der Unterschied zwischen den Bedingungen mehr als fünf Teilnehmer, so wurde der nächste Teilnehmer immer derjenigen Bedingung zugeordnet, die die geringste Anzahl Teilnehmer hatte.

9.2.10 Randomisierung der Antwortpositionen

Die Darbietung der Antwortoptionen eines Items, d.h. die Zuordnung zu den Antwortpositionen A, B und C, erfolgte randomisiert. Es gab sechs mögliche Permutationen für die drei Antwortoptionen: ABC, ACB, BAC, BCA, CAB, CBA. Jeder Permutation wurde eine Zahl zwischen eins und sechs zugeordnet. Vor der Darstellung des Items auf dem Bildschirm wurde durch Erzeugen einer Zufallszahl zwischen eins und sechs eine der Permutationen ausgewählt und die Antwortoptionen in der jeweiligen Reihenfolge angezeigt.

9.2.11 Randomisierung der Itemreihenfolge

Die Items wurden im Test in aufsteigender Schwierigkeit dargeboten. Die Randomisierung der Reihenfolge der Items wurde deshalb nur innerhalb der sechs Items einer Schwierigkeitsstufe vorgenommen. Für jeden Teilnehmer wurde zu Beginn des Tests eine zufällige individuelle Reihenfolge aller 36 Items generiert.

9.2.12 Auswahl von Teilnehmern

Das Experiment stand allen Nutzern des Internets zur Verfügung. In die Auswertung wurden jedoch nur die Daten von Teilnehmern einbezogen, wenn diese die folgenden Kriterien erfüllten:

- ***Muttersprache Deutsch:*** Um sicherzustellen, dass nur Testteilnehmer mit dem gleichen Idiom in die Stichprobe einbezogen wurden, wurden nur Teilnehmer berücksichtigt, die als Muttersprache „Deutsch“ ausgewählt hatten.
- ***Alter mindestens zehn Jahre:*** Den Begriff „Prozent“ verstehen Kinder erst ab ca. dem zehnten Lebensjahr. Es wurden deshalb keine Daten von Teilnehmern ausgewertet, die ein Alter geringer als zehn Jahre angegeben hatten.
- ***Vollständige Bearbeitung des Tests:*** Daten von Teilnehmern, die das Experiment nicht vollständig bearbeitet hatten, wurden ebenfalls nicht ausgewertet.
- ***Erstteilnahme:*** Um durch Mehrfachteilnahmen erzeugte Daten aus der Auswertung auszuschließen, wurde die Kombination aus IP-Adresse und Browserkennung überprüft. Wenn ein Teilnehmer mit einer bereits gespeicherten Kombination aus IP-Adresse und Browserkennung teilnahm, wurde er als „ungültig“ markiert und seine Daten aus der Auswertung ausgeschlossen. Dieses Kriterium allein kann eine Doppelteilnahme jedoch nicht sicher anzeigen, da es möglich ist, dass Teilnehmer von ihrem *Internet Service Provider* bei jedem neuen Zugriff auf das Internet eine andere IP-Adresse zugewiesen bekommen. Deshalb wurde dem Fragebogen (Testteil zwei) die folgende Frage zur Selbstauskunft des Teilnehmers vorangestellt: „Nehmen Sie an diesem Test zum ersten Mal teil?“. Beantwortete ein Teilnehmer diese Frage mit „nein“, wurden seine Daten ebenfalls nicht ausgewertet.

Der Ausschluss der Daten von Teilnehmern aufgrund der beschriebenen Kriterien wurde automatisiert durch die Applikation durchgeführt. So konnte im laufenden Experiment zu jeder Zeit eine annähernd gleich hohe Anzahl von Teilnehmern in allen Versuchsbedingungen sichergestellt werden.

9.2.13 Beschreibung der Stichprobe

Im Folgenden wird Multiple-Choice durch „MC“ abgekürzt und Multiple Evaluation durch „ME“. Die Stichprobe des ersten Experiments bestand aus 4824 Teilnehmern. Jede der sechs Versuchsbedingungen beinhaltete 804 Teilnehmer.

Die Altersverteilung in den Bedingungen lag im Durchschnitt bei 28,0 Jahren (*SD*=11,6; *SE*=0,167). Der jüngste Teilnehmer war 10 Jahre alt, da Teilnehmer, die jünger als 10 Jahre waren, aus der Auswertung ausgeschlossen wurden, und der älteste 88 Jahre alt. Tabelle 9-5 zeigt die deskriptive Statistik der Altersverteilung.

Tabelle 9-5

Deskriptive Statistik des Alters in den Bedingungen.

Bedingung	Minimum	Maximum	**Mittelwert**	Standardabweichung	Standardfehler	Median	N
ME-Dreieck	10	73	**27,8**	11,4	0,285	24	1608
ME-Schieberegler	12	88	**27,5**	11,2	0,280	24	1608
MC	11	71	**28,6**	12,2	0,305	24	1608
ME-ohne Auszahlungsinformation	10	70	**27,7**	10,8	0,231	24	2412
ME-mit Auszahlungsinformation	11	88	**28,2**	11,7	0,243	24	2412
ME-Dreieck ohne Auszahlungsinformation	10	65	**27,6**	10,9	0,386	24	804
ME-Dreieck mit Auszahlungsinformation	12	73	**28,0**	11,9	0,420	24	804
ME-Schieberegler ohne Auszahlungsinformation	12	67	**27,1**	10,8	0,379	24	804
ME-Schieberegler mit Auszahlungsinformation	12	88	**27,9**	11,7	0,411	24	804
MC-ohne Auszahlungsinformation	11	70	**28,3**	12,2	0,430	24	804
MC-mit Auszahlungsinformation	11	71	**28,9**	12,3	0,433	25	804

Am Experiment nahmen insgesamt 68,7% Frauen (*SD*=1,4%) und 31,3% Männer (*SD*=1,4%) teil. Tabelle 9-6 zeigt die Verteilung der Geschlechter in den Bedingungen.

Tabelle 9-6

Verteilung der Geschlechter in Prozent in den Versuchsbedingungen.

Bedingung	weiblich	männlich	N
ME-Dreieck	31,6%	68,4%	1608
ME-Schieberegler	29,8%	70,2%	1608
MC	32,6%	67,4%	1608
ME-ohne Auszahlungsinformation	30,8%	69,2%	2412
ME-mit Auszahlungsinformation	31,7%	68,3%	2412
ME-Dreieck-ohne Auszahlungsinformation	30,7%	69,3%	804
ME-Dreieck-mit Auszahlungsinformation	32,5%	67,5%	804
ME-Schieberegler-ohne Auszahlungsinformation	32,8%	67,2%	804
ME-Schieberegler-mit Auszahlungsinformation	32,0%	68,0%	804
MC-ohne Auszahlungsinformation	29,0%	71,0%	804
MC-mit Auszahlungsinformation	30,6%	69,4%	804

9.2.14 Rekrutierung der Teilnehmer

Die Teilnehmer des ersten Experiments wurden auf den in Tabelle 9-7 aufgeführten Webseiten auf das Experiment hingewiesen. Die Tabelle zeigt nur Webseiten, die mehr als 1% der Teilnehmer zum Test leiteten. Alle anderen wurden unter „Sonstige“ zusammengefasst.

Tabelle 9-7

Teilnehmer in Prozent, die von den Webseiten (Referrer) zum Englischtest geleitet wurden.

Internetadresse	Teilnehmer in Prozent
http://www.testedich.de/	19,4%
http://www.google.de/ und google.com/	19,2%
Sonstige	17,4%
http://www.englisch-lernen-im-internet.de/	16,7%
http://www.uni-duesseldorf.de/	11,4%
http://www.lernen.sprachdirekt.de/	7,2%
http://www.zv.uni-wuerzburg.de/	5,7%
http://idw-online.de/	1,8%
http://www.derberufsberater.de/	1,2%

9.3 Ergebnisse

9.3.1 Schwierigkeitsstufen

In den Bedingungen mit Multiple-Choice wählten die Testteilnehmer eine Antwort aus drei möglichen Optionen aus. In den Bedingungen mit Multipler Evaluation gaben die Teilnehmer hingegen ihre prozentuale Antwortsicherheit an. Die Schwierigkeit der Items wurde deshalb für die Bedingungen mit Multipler Evaluation und Multiple-Choice getrennt ermittelt. Je Schwierigkeitsstufe wurden jeweils zwölf Items gemeinsam betrachtet. Tabelle 9-8 zeigt die deskriptive Statistik der prozentualen Antwortsicherheit in die richtige Antwort, ermittelt über alle Bedingungen mit Multipler Evaluation.

Tabelle 9-8

Prozentuale Antwortsicherheit in die richtige Antwort, für einfache, mittelschwierige und schwierige Items für die Bedingungen mit Multipler Evaluation.

GER-Stufe	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	**Mittelwert/ Item**	Anzahl Items	N
A	einfach	168	1200	1047,21	178,52	3,15	1100,00	**87,27**	12	3216
B	mittel	97	1200	755,96	246,70	4,35	755,00	**63,00**	12	3216
C	schwierig	20	1200	533,05	212,46	3,75	479,00	**44,42**	12	3216
alle	alle	788	3600	2336,22	536,00	9,45	2326,50	**65,90**	36	3216

In jeder Schwierigkeitsstufe konnten die Teilnehmer in der Summe maximal 1200% Antwortsicherheit erreichen, insgesamt maximal 3600%.

Den richtigen Antwortoptionen der einfachen Items (Schwierigkeitsstufe A) ordneten die Testteilnehmer durchschnittlich 87,27% Antwortsicherheit je Item zu. Bei den mittelschwierigen Items (Schwierigkeitsstufe B) betrug die durchschnittliche prozentuale Antwortsicherheit in die richtige Antwort je Item 63,00% und bei den schwierigen Items (Schwierigkeitsstufe C) 44,42%.

In den Bedingungen mit Multiple-Choice betrug der mittlere Schwierigkeitsindex je Item (Bühner, 2006) für einfache Items 87,02%, für mittelschwierige Items 62,83% und für schwierige Items 44,16%. Tabelle 9-9 zeigt die deskriptive Statistik der richtigen Antworten in Prozent für die Bedingungen mit Multiple-Choice.

Tabelle 9-9

Richtige Antworten in Prozent, für einfache, mittelschwierige und schwierige Items für die Bedingungen mit Multiple-Choice.

GER-Stufe	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	**Mittelwert/ Item**	Anzahl Items	N
A	einfach	200	1200	1044,22	192,72	4,81	1100,00	**87,02**	12	1608
B	mittel	100	1200	753,98	265,61	6,62	800,00	**62,83**	12	1608
C	schwierig	0	1200	539,91	249,07	6,21	500,00	**44,16**	12	1608
alle	alle	600	3600	2333,52	551,50	7,94	2308,00	**64,82**	36	1608

In jeder Schwierigkeitsstufe konnten die Teilnehmer in der Summe maximal 1200% richtige Antworten erreichen, insgesamt maximal 3600%.

9.3.2 Punktauszahlungen

Zur Ermittlung der Punktsumme eines Teilnehmers wurden die Punktauszahlungen aller 36 Items summiert. Zuerst werden die Punktsummen der Bedingungen mit Multipler Evaluation betrachtet. Der Auszahlungsbereich der Auswertefunktion lag hierbei zwischen -100 und 100 Punkten je Item. Die maximal erreichbare Punktsumme betrug also 3600 Punkte. Tabelle 9-10 zeigt die deskriptive Statistik der Punktsummen. Gaben die Teilnehmer ihre Antwortsicherheiten durch die Schieberegler wieder, so erzielten sie mit 1425,80 Punkten durchschnittlich 38,80 Punkte mehr (je Item durchschnittlich 1,08 Punkte) als die Teilnehmer, die das Dreieck verwendeten und damit 1387,0 Punkte erreichten. Diese Punktedifferenz war jedoch in einer zweifaktoriellen Varianzanalyse über die Punktsummen mit den Faktoren „Antwortinstrument“ und „Auszahlungsinformation“ für den Faktor „Antwortinstrument“ nicht signifikant (F(1, 3215)=1,319; p=,251; η^2<,001). Erhielten die Testteilnehmer eine Auszahlungsinformation, erreichten sie mit 1456,50 Punkten im Durchschnitt 100,20 Punkte (je Item durchschnittlich 2,80 Punkte) mehr als die Teilnehmer, die keine Auszahlungsinformation erhielten und 1356,30 Punkte erzielten. Für den Faktor „Auszahlungsinformation“ war die Punktedifferenz signifikant (F(1, 3215)=8,815; p=,003; η^2=,003), jedoch war die Effektstärke gering. Eine signifikante Interaktion zwischen dem verwendeten Antwortinstrument und ob die Teilnehmer die Information über die Auszahlung erhielten, belegte die zweifaktorielle Varianzanalyse nicht (F(1, 3215)=1,544; p=,214; η^2<,001).

Tabelle 9-10

Deskriptive Statistik der Punktsummen aller Items des Tests für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	**Mittelwert**	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
ME-Dreieck	-1700	3600	**1387,00**	979,12	24,40	1390,00	38,50	1608
ME-Schieberegler	-1396	3600	**1425,80**	937,81	23,40	1424,00	39,60	1608
ME-ohne Auszahlungsinformation	-1323	3600	**1356,30**	973,80	24,30	1352,00	37,70	1608
ME-mit Auszahlungsinformation	-1700	3600	**1456,50**	941,06	23,50	1466,50	40,50	1608
ME-Dreieck-ohne Auszahlungsinformation	-1323	3600	**1357,87**	996,31	35,14	1337,50	37,72	804
ME-Dreieck-mit Auszahlungsinformation	-1700	3600	**1416,17**	961,35	33,90	1443,00	39,34	804
ME-Schieberegler-ohne Auszahlungsinformation	-1315	3600	**1354,70**	951,38	33,55	1365,50	37,63	804
ME-Schieberegler-mit Auszahlungsinformation	-1396	3600	**1496,91**	919,14	32,42	1494,00	41,58	804

Anzahl Items=36.

Wurden die Punktsummen für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items getrennt betrachtet, so verringerte sich, wie zu erwarten war, die mittlere Punktsumme in Abhängigkeit von der Schwierigkeit der Items. Eine dreifaktorielle Varianzanalyse über die Faktoren „Antwortinstrument", „Auszahlungsinformation" und einer Messwiederholung auf dem dritten Faktor „Schwierigkeit" zeigte nach einer Greenhouse-Geisser-Korrektur, dass dieser Einfluss der Schwierigkeit auf die Punktsummen deutlich signifikant war (F(1.978, 6351.947))=7929,775; p<,001; η^2=,712). Die Verwendung der Schieberegler führte zu einer etwas höheren Punktsumme der Testteilnehmer, besonders bei den mittelschwierigen und schwierigen Items im Vergleich zur Wiedergabe von Antwortsicherheiten mithilfe des Dreiecks. Abbildung 9-5 zeigt die Mittelwerte der Punktsummen für die Bedingungen „ME-Dreieck" und „ME-Schieberegler". Eine dreifaktorielle Varianzanalyse über die Faktoren „Antwortinstrument", „Auszahlungsinformation" und einer Messwiederholung auf dem dritten Faktor „Schwierigkeit" zeigte nach einer Greenhouse-Geisser-Korrektur jedoch keine

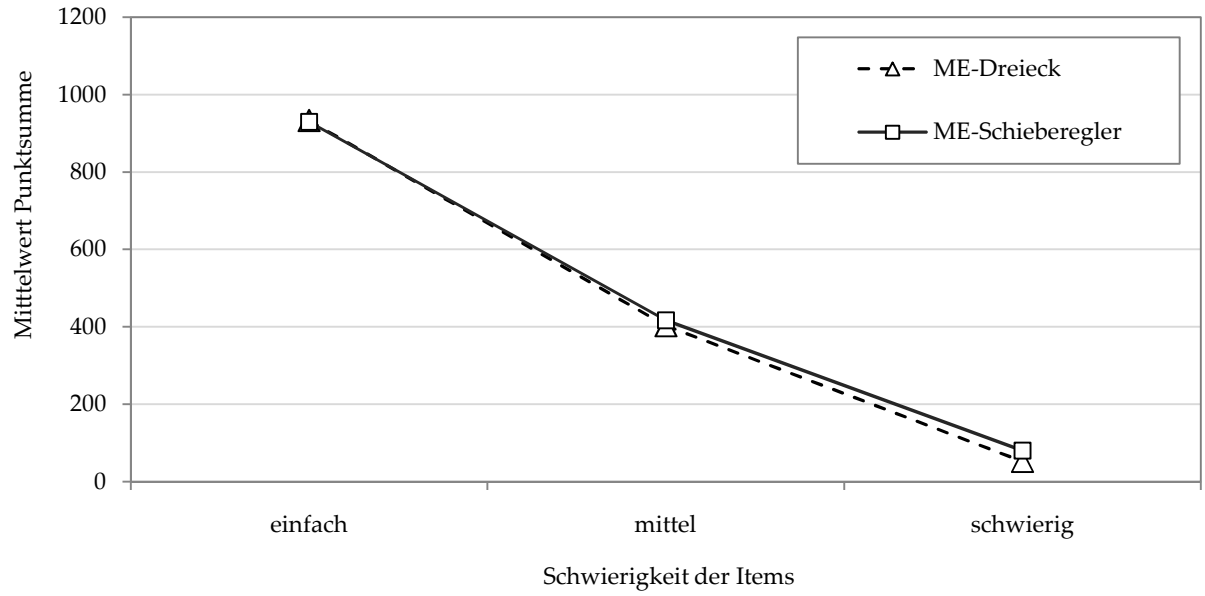

Abbildung 9-5: Mittelwerte der Punktsummen für die Bedingungen „ME-Dreieck“ und „ME-Schieberegler“. Maximal konnten die Teilnehmer bei je zwölf Items pro Schwierigkeitsstufe 1200 Punkte erzielen. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

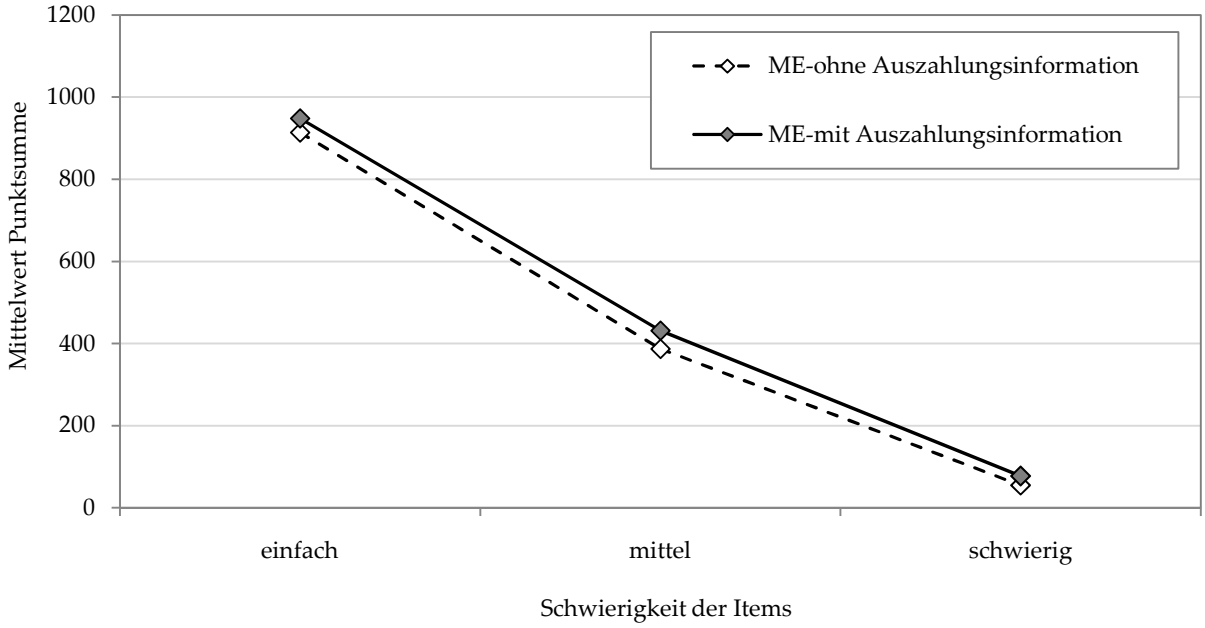

Abbildung 9-6: Mittelwerte der Punktsummen für die Bedingungen mit Multipler Evaluation mit und ohne Auszahlungsinformation. Maximal konnten die Teilnehmer bei je zwölf Items pro Schwierigkeitsstufe 1200 Punkte erzielen. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

signifikante Interaktion zwischen der Schwierigkeit der Items und dem Antwortinstrument (F(1.978, 6351.947)=7929,775; p=,085; η^2=,001). Abbildung 9-6 zeigt die Mittelwerte der Punktsummen für die Bedingungen mit und ohne Auszahlungsinformation ebenfalls für einfache, mittelschwierige und schwierige Items. Auf allen Schwierigkeitsstufen erzielten die Testteilnehmer, die eine Auszahlungsinformation erhielten, eine etwas höhere Punktsumme. Eine signifikante Interaktion zwischen der Schwierigkeit der Items und ob die Teilnehmer eine Auszahlungsinformation erhielten bestand jedoch nicht. Dies zeigte eine dreifaktorielle Varianzanalyse über die Faktoren „Antwortinstrument", „Auszahlungsinformation" und einer Messwiederholung auf dem dritten Faktor „Schwierigkeit" nach einer Greenhouse-Geisser-Korrektur (F(1.978, 6351.947)=1,284; p=,277; η^2<,001). Eine signifikante Interaktion zwischen der Schwierigkeit der Items, dem Vorhandensein einer Auszahlungsinformation und der Art des verwendeten Antwortinstruments zeigte diese Varianzanalyse ebenfalls nicht (F(1.978, 6351.947)=1,434; p=,238; η^2<,001). Die Tabellen mit der deskriptiven Statistik der Punktsummen, getrennt für einfache, mittelschwierige und schwierige Items, und die Mittelwerte der Punktauszahlungen je Item befinden sich im Anhang A.1.

Im Folgenden werden die Punktsummen der Bedingungen mit Multiple-Choice betrachtet. Die Teilnehmer erhielten einen Punkt für eine richtige und null Punkte für eine falsche Antwort. Tabelle 9-11 zeigt die deskriptive Statistik der Punktsummen, ermittelt für alle 36 Items des Englischtests. Die Teilnehmer erzielten

Tabelle 9-11

Deskriptive Statistik der Punktsummen aller 36 Items für die Bedingungen mit Multiple-Choice.

Bedingung	Minimum	Maximum	**Mittelwert**	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
MC- ohne Auszahlungsinformation	6	36	**22,96**	5,84	0,206	23,00	0,638	804
MC-mit Auszahlungsinformation	6	36	**23,60**	5,78	0,204	23,00	0,656	804

mit 22,96 Punkten durchschnittlich 0,64 Punkte mehr, wenn sie eine Auszahlungsinformation erhielten. Dieser Mittelwertunterschied zeigte sich in einer einfaktoriellen Varianzanalyse über die Punktsummen mit dem Faktor „Auszahlungsinformation" als signifikant (F(1, 1607)=4,834; p=,028; $\eta^{2=}$,003).

Auch für die Bedingungen mit Multiple-Choice zeigte sich in einer zweifaktoriellen Varianzanalyse mit dem Faktor „Auszahlungsinformation“ und einer Messwiederholung auf dem zweiten Faktor „Schwierigkeit“ nach einer Greenhouse-Geisser-Korrektur ein deutlicher signifikanter Einfluss der Schwierigkeit auf die Punktsummen der Teilnehmer (F(1.954, 3137.586)=3746,386; p<,001; η^2=,700). Zwischen der Schwierigkeit der Items und ob die Teilnehmer eine Auszahlungsinformation erhielten bestand keine signifikante Interaktion (F(1.954, 3137.586)=0,248; p<,776; η^2<,001). Abbildung 9-7 zeigt die Mittelwerte der Punktsummen für die Bedingungen „MC-ohne Auszahlungsinformation“ und „MC-mit Auszahlungsinformation“ der einfachen, mittelschwierigen und schwierigen Items. Die Tabellen mit der deskriptiven Statistik der Punktsummen, getrennt für einfache, mittelschwierige und schwierige Items, und die Mittelwerte der Punktauszahlungen je Item befinden sich in Anhang A.1.

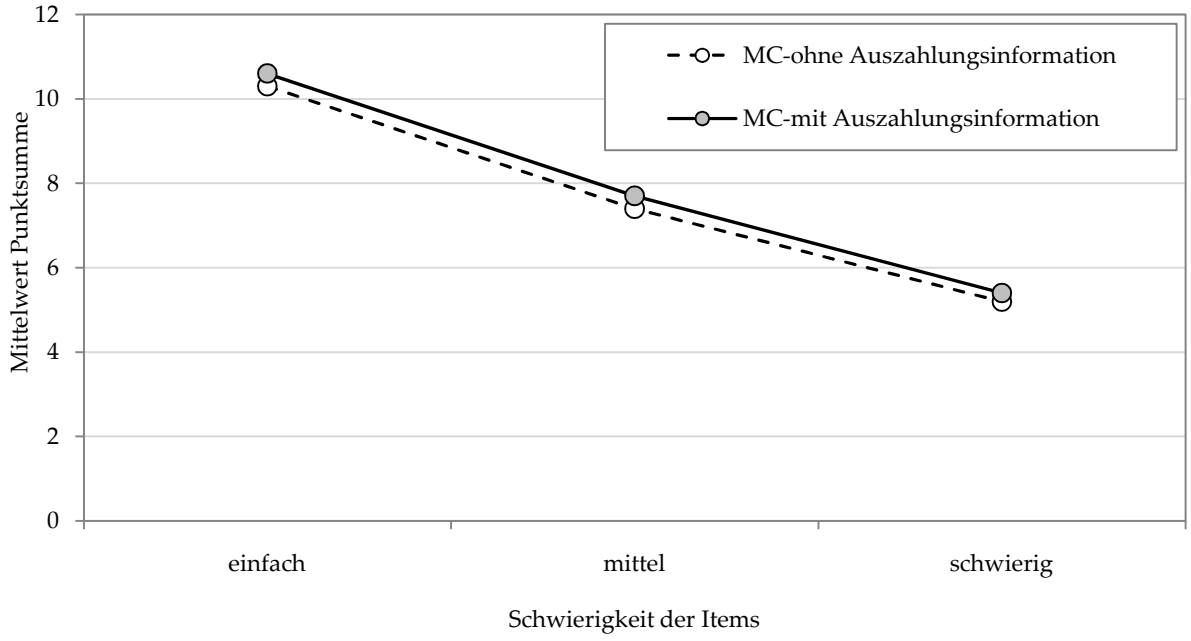

Abbildung 9-7: Mittelwerte der Punktsummen für die Bedingungen mit Multiple-Choice mit und ohne Auszahlungsinformation. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

9.3.3 Trennschärfen

Für die Punktauszahlungen je Item wurden *part-whole-korrigierte* Trennschärfen ermittelt. Die niedrigste Trennschärfe betrug 0,06 für ein schwieriges Item in der Bedingung „MC-ohne Auszahlungsinformation“, die höchste 0,46 für ein mittelschwieriges Item in der Bedingung „ME-Schieberegler mit Auszahlungsinformation“. Die *part-whole-korrigierten* Trennschärfen der einzelnen Items sind in Anhang A.2 aufgeführt. Zur Berechnung der durchschnittlichen Trennschärfe pro Bedingung wurden die Trennschärfen Fisher-Z-transformiert und die Mittelwerte dieser Z-transformierten Trennschärfen gebildet. Die Mittelwerte wurden einer inversen Fisher-Z-Transformation unterzogen. Die Bedingungen mit Multipler Evaluation zeigten eine durchschnittliche Trennschärfe von 0,33 und die Bedingungen mit Multiple-Choice von 0,31. Die Trennschärfen der Bedingungen mit Multipler Evaluation waren also etwas höher. Tabelle 9-12 zeigt die Mittelwerte und Standardabweichungen der *part-whole-korrigierten* Trennschärfen.

Tabelle 9-12

Mittelwerte und Standardabweichungen der *part-whole-korrigierten* Trennschärfen.

Bedingung	**Mittelwert**	Standardabweichung	Anzahl Items	N
ME	**0,33**	0,09	36	3216
MC	**0,31**	0,11	36	1608
ME-Dreieck	**0,33**	0,08	36	1608
ME-Schieberegler	**0,33**	0,09	36	1608
ME-ohne Auszahlungsinformation	**0,33**	0,09	36	1608
ME-mit Auszahlungsinformation	**0,33**	0,09	36	1608
ME-Dreieck-ohne Auszahlungsinformation	**0,34**	0,08	36	804
ME-Dreieck-mit Auszahlungsinformation	**0,32**	0,08	36	804
ME-Schieberegler-ohne Auszahlungsinformation	**0,33**	0,10	36	804
ME-Schieberegler-mit Auszahlungsinformation	**0,34**	0,10	36	804
MC-ohne Auszahlungsinformation	**0,31**	0,11	36	804
MC-mit Auszahlungsinformation	**0,31**	0,11	36	804

Zwischen den Versuchsbedingungen mit Multipler Evaluation zeigte sich kein Unterschied zwischen den Trennschärfen in Abhängigkeit davon, mit welchem Antwortinstrument die Antwortsicherheiten erhoben wurden, oder ob die

Testteilnehmer eine Auszahlungsinformation erhielten. Der Mittelwert der *part-whole-korrigierten* Trennschärfen betrug jeweils 0,33. Auch für die Bedingungen mit Multiple-Choice zeigte sich mit einer mittleren Trennschärfe von 0,31 jeweils kein Unterschied zwischen den Bedingungen mit und ohne Auszahlungsinformation. Tabelle 9-13 zeigt die Trennschärfen, getrennt für einfache, mittelschwierige und schwierige Items, für die Versuchsbedingungen. Der Test mit Multipler Evaluation zeigte für alle drei Schwierigkeitsstufen eine höhere Trennschärfe gegenüber dem Test mit Multiple-Choice, und zwar umso deutlicher, je höher die Schwierigkeitsstufen der Items waren. Die höchste Verbesserung von 0,25 auf 0,28 um 0,03 zeigte sich für schwierige Items. Wurden die Schieberegler zur Wiedergabe der Antwortsicherheit verwendet, so waren die Mittelwerte der Trennschärfen für einfache und mittelschwierige Items etwas höher als wenn das Dreieck verwendet wurde. Keine Unterschiede zeigten sich für die Trennschärfen in Abhängigkeit davon, ob die Testteilnehmer eine Auszahlungsinformation erhielten oder nicht.

Tabelle 9-13

Mittelwerte und Standardabweichungen der *part-whole-korrigierten* Trennschärfen.

Schwierigkeit	einfach		mittel		schwierig		
Bedingung	**Mittelwert**	Standardabweichung	**Mittelwert**	Standardabweichung	**Mittelwert**	Standardabweichung	N
ME	**0,33**	0,08	**0,36**	0,11	**0,29**	0,05	3216
MC	**0,32**	0,08	**0,35**	0,11	**0,26**	0,10	1608
ME-Dreieck	**0,32**	0,07	**0,35**	0,10	**0,30**	0,05	1608
ME-Schieberegler	**0,34**	0,09	**0,36**	0,11	**0,29**	0,06	1608
ME-ohne Auszahlungsinformation	**0,33**	0,07	**0,36**	0,11	**0,30**	0,05	1608
ME-mit Auszahlungsinformation	**0,33**	0,08	**0,36**	0,10	**0,29**	0,06	1608
ME-Dreieck-ohne Auszahlungsinformation	**0,32**	0,07	**0,37**	0,11	**0,32**	0,06	804
ME-Dreieck-mit Auszahlungsinformation	**0,32**	0,07	**0,35**	0,10	**0,28**	0,06	804
ME-Schieberegler-ohne Auszahlungsinformation	**0,34**	0,08	**0,35**	0,12	**0,28**	0,06	804
ME-Schieberegler- mit Auszahlungsinformation	**0,34**	0,10	**0,38**	0,11	**0,30**	0,07	804
MC-ohne Auszahlungsinformation	**0,32**	0,08	**0,35**	0,12	**0,25**	0,10	804
MC-mit Auszahlungsinformation	**0,31**	0,08	**0,35**	0,12	**0,28**	0,11	804

9.3.4 Reliabilität

Zur Bestimmung der internen Konsistenz des Englischtests wurde der Cronbach-α-Koeffizient (Cronbach, 1951) als Maß der Reliabilität für die Punktauszahlung je Item ermittelt. Die Prüfung auf eine Signifikanz der Unterschiede zwischen den Cronbach-α-Koeffizienten der Versuchsbedingungen wurde mit dem Programm *alphatst.exe* (Lautenschlager, 1989) nach der Methode von Feldt, Woodruff und Salih (1987) durchgeführt.

Zunächst werden die Cronbach-α-Koeffizienten, ermittelt für alle 36 Items des Tests, betrachtet. Es wurde geprüft, ob die Bedingungen mit Multipler Evaluation gegenüber den Bedingungen mit Multiple-Choice eine verbesserte Reliabilität zeigten. Wurden alle Bedingungen mit Multipler Evaluation zusammengefasst (N=3216), so betrug α 0,84. Damit zeigte sich bei der Durchführung des Tests mit Multipler Evaluation eine signifikante Verbesserung der Reliabilität um 0,02 gegenüber der Durchführung des Tests mit Multiple-Choice (N=1608). Dort betrug der Cronbach-α-Koeffizient nur 0,82 (X^2=7,115; *df*=1; *p*=,008).

Tabelle 9-14

Cronbach-α-Koeffizienten, ermittelt für die Punktauszahlung aller 36 Items des Tests.

Bedingung	**α**	α für standardisierte Items	N
ME	**0,84**	0,84	3216
MC	**0,82**	0,83	1608
ME-Dreieck	**0,84**	0,84	1608
ME-Schieberegler	**0,84**	0,84	1608
ME-ohne Auszahlungsinformation	**0,84**	0,84	1608
ME-mit Auszahlungsinformation	**0,84**	0,84	1608
ME-Dreieck-ohne Auszahlungsinformation	**0,84**	0,85	804
ME-Dreieck-mit Auszahlungsinformation	**0,83**	0,83	804
ME-Schieberegler-ohne Auszahlungsinformation	**0,84**	0,84	804
ME-Schieberegler-mit Auszahlungsinformation	**0,84**	0,85	804
MC-ohne Auszahlungsinformation	**0,82**	0,82	804
MC-mit Auszahlungsinformation	**0,82**	0,83	804

Wurden die Antwortsicherheiten mithilfe der Schieberegler erhoben, betrug α 0,84 und war gegenüber der Erhebung durch das Dreieck, mit ebenfalls 0,84, nicht

verbessert. Tabelle 9-14 zeigt die Cronbach-α-Koeffizienten für die Versuchsbedingungen. Nur die Bedingung mit dem Antwortdreieck und einer Auszahlungsinformation zeigte mit einem α von 0,83 eine etwas geringere Reliabilität. Dieser Unterschied war im Vergleich zu den Bedingungen mit den Schiebereglern und einem α von 0,84 jedoch nicht signifikant (X^2=0,1378; *df*=1; *p*=,674). Ebenso zeigte sich kein signifikanter Unterschied in der Reliabilität in Abhängigkeit davon, ob die Testteilnehmer eine Auszahlungsinformation erhielten oder nicht. Für beide Bedingungen betrug die interne Konsistenz 0,84. Für die Bedingungen mit Multiple-Choice zeigte sich ebenfalls keine Verbesserung der Reliabilität, wenn die Testteilnehmer eine Information über die Auszahlung erhielten.

Im Folgenden werden die Cronbach-α-Koeffizienten, ermittelt für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items, betrachtet. Da nur

Tabelle 9-15

Cronbach-α-Koeffizienten ermittelt für die Punktauszahlung der jeweils zwölf einfachen, mittelschwierigen und schwierigen Items.

	α			α für standardisierte Items			
Bedingung/Schwierigkeit	einfach	mittel	schwierig	einfach	mittel	schwierig	N
ME	**0,72**	**0,69**	**0,66**	0,73	0,69	0,66	3216
MC	**0,71**	**0,68**	**0,58**	0,72	0,68	0,58	1608
ME-Dreieck	**0,71**	**0,69**	**0,66**	0,72	0,69	0,66	1608
ME-Schieberegler	**0,73**	**0,70**	**0,65**	0,74	0,70	0,66	1608
ME-ohne Auszahlungsinformation	**0,72**	**0,70**	**0,67**	0,72	0,69	0,67	1608
ME-mit Auszahlungsinformation	**0,72**	**0,69**	**0,64**	0,73	0,69	0,64	1608
ME-Dreieck-ohne Auszahlungsinformation	**0,71**	**0,71**	**0,69**	0,72	0,71	0,69	804
ME-Dreieck-mit Auszahlungsinformation	**0,70**	**0,67**	**0,63**	0,72	0,68	0,63	804
ME-Schieberegler-ohne Auszahlungsinformation	**0,72**	**0,68**	**0,66**	0,73	0,68	0,66	804
ME-Schieberegler-mit Auszahlungsinformation	**0,73**	**0,71**	**0,65**	0,75	0,71	0,66	804
MC-ohne Auszahlungsinformation	**0,70**	**0,67**	**0,58**	0,71	0,67	0,58	804
MC-mit Auszahlungsinformation	**0,71**	**0,68**	**0,60**	0,72	0,68	0,60	804

jeweils zwölf Items berücksichtigt wurden, war die Reliabilität niedriger als die des gesamten Tests. Beide Antwortverfahren zeigten zudem mit zunehmender

Schwierigkeit der Items eine Verschlechterung der Reliabilität. Tabelle 9-15 zeigt die Cronbach-α-Koeffizienten für die Items der drei Schwierigkeitsstufen. Wurden alle Bedingungen mit Multipler Evaluation zusammengefasst, so betrug α für einfache Items 0,72. Damit wurde für den Test mit Multipler Evaluation bei einfachen Items eine nicht signifikante Verbesserung der Reliabilität um 0,01 gegenüber dem Test mit Multiple-Choice mit einem α von 0,71 (X^2=0,5603; *df*=1; *p*=,461) gezeigt. Auch für die mittelschwierigen Items war die Reliabilität der Bedingungen mit Multipler Evaluation mit einem Cronbach-α von 0,69 nicht signifikant um 0,01 im Vergleich zu den Bedingungen mit Multiple-Choice verbessert (X^2=0,4585; *df*=1; *p*=,506). Eine mit einem α von 0,66 um 0,08 signifikant höhere Reliabilität zeigte sich für schwierige Items in den Bedingungen mit Multipler Evaluation, gegenüber einer Reliabilität von 0,58 in den Bedingungen mit Multiple-Choice (X^2=20,6821; *df*=1; *p*<,001).

Für einfache, mittelschwierige und schwierige Items zeigten sich keine signifikanten Unterschiede zwischen den Bedingungen mit Multipler Evaluation in Abhängigkeit davon, ob die Testteilnehmer während des Experiments die Schieberegler oder das Antwortdreieck verwendet hatten. Tabelle 9-16 zeigt die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Bedingungen.

Tabelle 9-16

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten der Bedingungen für den Faktor „Antwortinstrument".

Schwierigkeit	Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
einfach		0,71		0,73	**0,02**	1,735	1	,185
mittel	ME-Dreieck	0,69	ME-Schieberegler	0,70	**0,01**	0,365	1	,552
schwierig		0,66		0,65	**-0,01**	0,286	1	,595

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=α 2-α 1, positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, N je Bedingung=1608.

Ebenfalls zeigten sich keine signifikanten Unterschiede für die Reliabilität, ermittelt für die drei Schwierigkeitsstufen, in Abhängigkeit davon, ob die Testteilnehmer eine Auszahlungsinformation erhielten oder nicht, sowohl für die Bedingungen mit Multipler Evaluation, als auch für die Bedingungen mit Multiple-Choice. Tabelle 9-17 zeigt die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Bedingungen.

Tabelle 9-17

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten der Bedingungen für den Faktor „Auszahlungsinformation“.

Schwierigkeit	Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\,\alpha$**	X^2	*df*	*p*
einfach	ME-ohne Auszahlungsinformation	0,72	ME-mit Auszahlungsinformation	0,72	**0**	-	-	-
mittel		0,70		0,69	**-0,01**	0,365	1	,552
schwierig		0,67		0,64	**-0,03**	2,571	1	,105
einfach	MC-ohne Auszahlungsinformation	0,70	MC-mit Auszahlungsinformation	0,71	**0,01**	0,404	1	,532
mittel		0,67		0,68	**0,01**	0,195	1	,645
schwierig		0,58		0,60	**0,02**	0,161	1	,663

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2 = Cronbachs-α der zweiten Bedingung, $\Delta\,\alpha$=α 2-α 1, positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung,
N=1608 für die Bedingungen mit ME und N=804 für die Bedingungen mit MC.

Zusammengefasst zeigten die Ergebnisse des Experiments eine signifikant höhere interne Konsistenz des Tests mit Multipler Evaluation, insbesondere bei schwierigen Items, im Vergleich zu dem Test mit Multiple-Choice. Die prozentgenaue Erhebung von Antwortsicherheit mithilfe der Schieberegler führte nicht zu einer Verbesserung der Reliabilität im Vergleich zur Erhebung mit dem Antwortdreieck mit nur 16 Antwortmöglichkeiten. Eine Information der Teilnehmer über die Auszahlung bewirkte keine Verbesserung der Reliabilität des Tests mit Multipler Evaluation, verglichen mit den Bedingungen, in denen die Teilnehmer diese Information nicht erhielten.

9.3.5 Realismusindex

Mithilfe des individuellen Realismusindex kann die Güte der Kalibrierung eines Teilnehmers bewertet werden. Ein perfekt kalibrierter Testteilnehmer erzielt dabei einen Realismusindex a=1. Bei einem Teilnehmer, der *Overconfidence* zeigt, ist a<1. Bei einem, der *Underconfidence* zeigt, ist a>1. Wurden alle Teilnehmer der Bedingungen mit Multipler Evaluation betrachtet, so zeigte sich ein durchschnittlicher Realismusindex von a=0,64 (*SD*=0,24; *SE*=0,0042). Die Teilnehmer zeigten also im Mittel *Overconfidence*. Der niedrigste ermittelte Realismusindex betrug dabei -0,51,

der höchste 1,15. Tabelle 9-18 zeigt die deskriptive Statistik des Realismusindex der Teilnehmer für alle Bedingungen mit Multipler Evaluation. Abbildung 9-8 zeigt

Tabelle 9-18

Deskriptive Statistik des Realismusindex der Testteilnehmer.

Bedingung	Minimum	Maximum	**Mittelwert**	Standardfehler	Standardabweichung	Median	Schiefe	N
ME-Dreieck	-0,28	1,06	**0,62**	0,006	0,24	0,65	-0,60	1608
ME-Schieberegeler	-0,51	1,15	**0,66**	0,006	0,24	0,69	-0,74	1608
ME-ohne Auszahlungsinformation	-0,51	1,14	**0,62**	0,006	0,25	0,64	-0,54	1608
ME-mit Auszahlungsinformation	-0,37	1,15	**0,65**	0,006	0,23	0,69	-0,80	1608
ME-Dreieck ohne Auszahlungsinformation	-0,23	1,06	**0,62**	0,009	0,25	0,64	-0,50	804
ME-Dreieck mit Auszahlungsinformation	-0,28	1,03	**0,62**	0,008	0,23	0,66	-0,73	804
ME-Schieberegler ohne Auszahlungsinformation	-0,51	1,14	**0,63**	0,009	0,25	0,65	-0,58	804
ME-Schieberegler mit Auszahlungsinformation	-0,37	1,15	**0,68**	0,008	0,23	0,72	-0,92	804

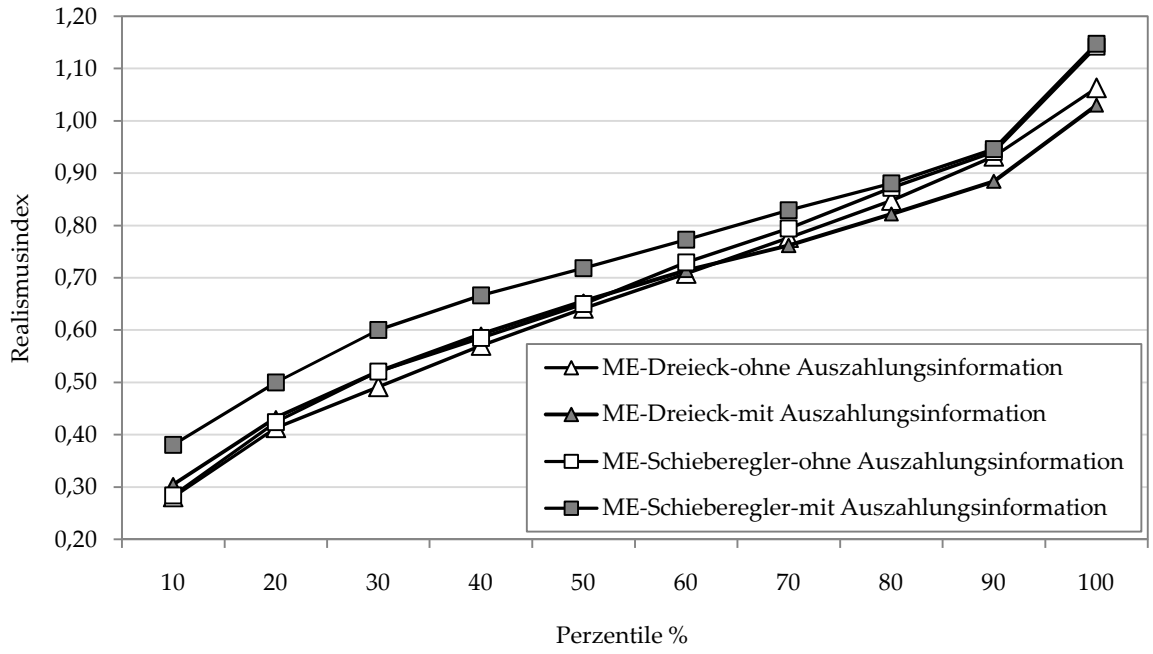

Abbildung 9-8: Realismusindex in den Versuchsbedingungen, dargestellt für zehn Perzentile.

zeigt den Realismusindex a für die Versuchsbedingungen, unterteilt in zehn Perzentile. Die Testteilnehmer zeigten mit a=0,66 einen um 0,04 signifikant höheren Realismusindex, wenn sie ihre Antwortsicherheiten mit den Schiebereglern angaben, im Vergleich zur Angabe mit dem Dreieck mit a=0,62. Dies zeigte eine zweifaktorielle Varianzanalyse über den Realismusindex mit den Faktoren „Antwortinstrument" und „Auszahlungsinformation" für den Faktor „Antwortinstrument" ($F(1, 3215)=22{,}655$; $p<{,}001$; $\eta^2={,}007$)). Die Bedingungen „mit Auszahlungsinformation" zeigten ebenfalls einen signifikant um 0,03 verbesserten durchschnittlichen Realismusindex von a=0,65 gegenüber den Bedingungen ohne Auszahlungsinformation mit a=0,62. Dies belegte die zweifaktorielle Varianzanalyse für den Faktor „Auszahlungsinformation" ($F(1, 3215)=10{,}562$; $p={,}001$; $\eta^2={,}003$). Zwischen den beiden Faktoren „Antwortinstrument" und „Auszahlungsinformation" bestand zudem eine signifikante Interaktion ($F(1, 3215)=8{,}491$; $p={,}004$; $\eta^2={,}003$). Den höchsten mittleren Realismusindex von a=0,68 zeigten die Testteilnehmer, wenn sie ihre Antwortsicherheiten mit den Schiebereglern angaben und eine Auszahlungsinformation erhielten. Den niedrigsten Mittelwert von 0,62 erreichten sie hingegen, wenn sie das Dreieck ohne Auszahlungsinformation verwendeten. Die Differenz zwischen beiden Bedingungen betrug 0,06.

9.3.6 Punktauszahlungen nach Realismuskorrektur

Bei der Realismuskorrektur werden die Antwortsicherheiten in die richtige Antwort für jeden Testteilnehmer auf der Basis seines individuellen Realismusindex korrigiert. Anhand dieser korrigierten Antwortsicherheiten wurden die Punkte erneut ausgezahlt. Im Folgenden wird die Korrektur zunächst für zwei einzelne Teilnehmer verdeutlicht. Der eine Teilnehmer zeigte *Overconfidence* und hatte einen Realismusindex von a=0,35. Der andere zeigte *Underconficence* und hatte einen Realismusindex von a=1,15. Abbildung 9-9 zeigt die nicht korrigierten Punktauszahlungen und die auf der Basis des Realismusindex korrigierten Auszahlungen für den Teilnehmer, der *Overconfidence* zeigte. Um die Korrektur grafisch zu veranschaulichen, werden alle 36 Auszahlungen nach ihrer Höhe sortiert dargestellt. Durch die Korrektur der Antwortsicherheiten erhielt der Teilnehmer die Antwortsicherheiten, die er hätte berichten müssen, wenn er perfekt kalibriert

gewesen wäre. Da der Teilnehmer eine deutliche *Overconfidence* zeigte, weichen die korrigierten Antwortsicherheiten und, wie Abbildung 9-9 zeigt, damit auch die erneut ausgezahlten Punkte stark von den nicht korrigierten ab. Gab der Teilnehmer beispielsweise 0% Antwortsicherheit in eine richtige Antwort wieder, so wurde diese Antwortsicherheit durch die Korrektur auf 22% erhöht und damit die Höhe der ausgezahlten Strafpunkte von -100 Punkte auf -27 Punkte reduziert. Im Gegensatz dazu wurden beispielsweise 100% Antwortsicherheit in eine richtige Antwortoption auf 56% reduziert und damit die dafür ausgezahlten Punkte von 100 auf 43 Punkte verringert. Gab der Teilnehmer mit 33% Antwortsicherheit völliges Nichtwissen an, so blieb diese Antwortsicherheit von 33% auch nach der Korrektur unverändert, und der Teilnehmer erhielt erneut eine Auszahlung von 0 Punkten. Insgesamt wurde die Punktsumme des Teilnehmers aufgrund der Korrektur auf der Basis des Realismusindex von 268 Punkten um 34 Punkte auf 302 Punkte erhöht.

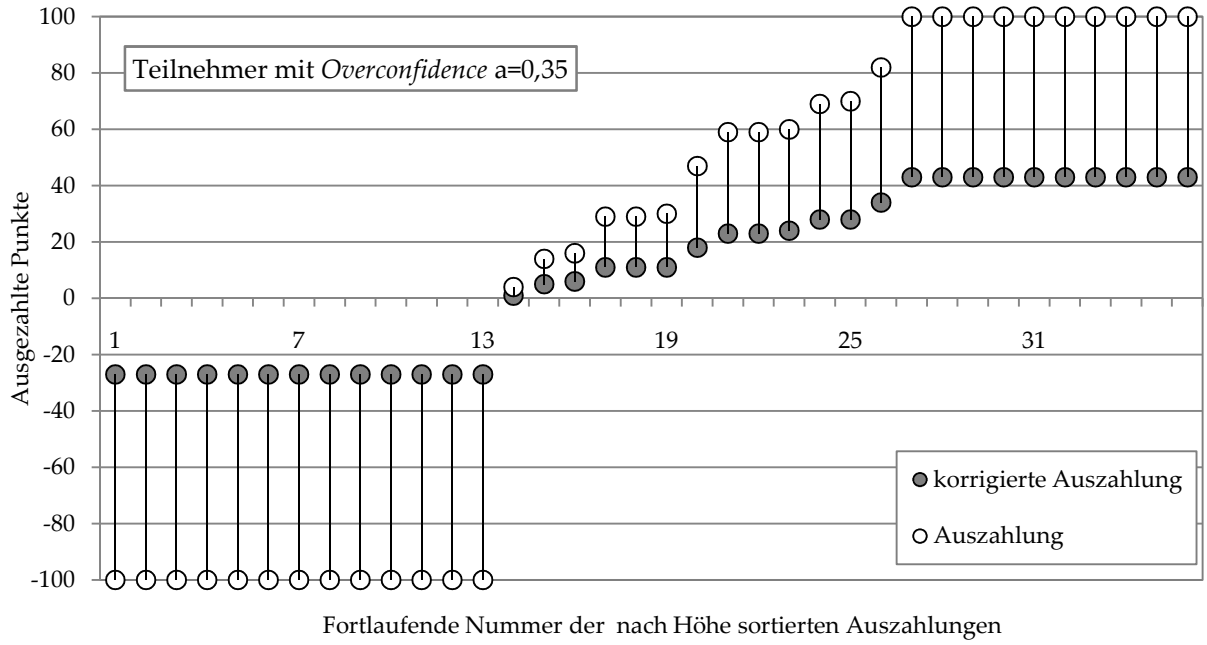

Abbildung 9-9: Korrigierte und nicht korrigierte Auszahlungen eines Teilnehmers mit *Overconfidence* und einem Realismusindex von a=0,35. Die Auszahlungen werden in nach Höhe sortierter Reihenfolge dargestellt.

Abbildung 9-10 zeigt das Beispiel eines Teilnehmers, der mit einem Realismusindex von a=1,15 eine geringfügige *Underconfidence* zeigte. Bei einem Teilnehmer mit *Underconfidence* wurde die in die richtige Antwort berichtete Antwortsicherheit in Abhängigkeit von seinem individuellen Realismusindex durch die Korrektur erhöht. Damit erhielt er für Antwortsicherheiten größer als 33% mehr Punkte ausgezahlt, für Antwortsicherheiten in die richtige Antwort, die kleiner als 33% waren, erhielt er jedoch auch höhere Strafzahlungen. In dem Beispiel eines Teilnehmers, der mit einem Realismusindex von a=1,15 *Underconfidence* zeigt, wurden, wie in Abbildung *9-10* dargestellt, die Auszahlungen geringfügiger korrigiert im Vergleich zu dem Teilnehmer mit *Overconfidence*, mit einem Realismusindex von a=0,35. Vor der Korrektur bekam der Teilnehmer insgesamt 631 Punkte ausgezahlt, nach der Korrektur 701 Punkte und damit 70 Punkte mehr.

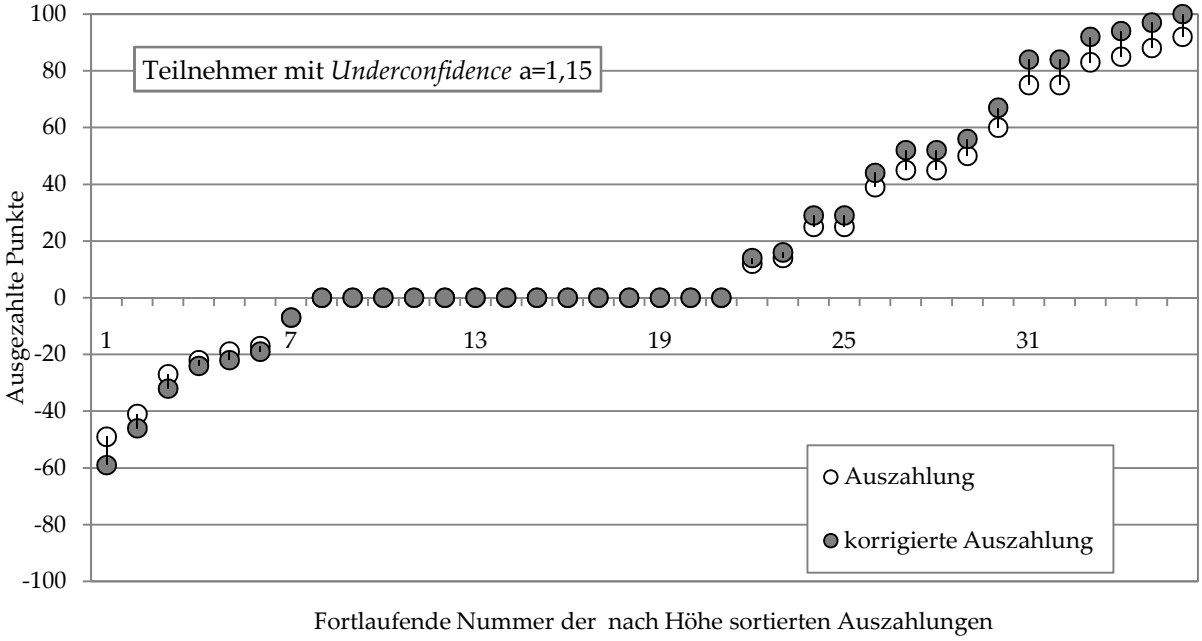

Abbildung 9-10: Korrigierte und nicht korrigierte Auszahlungen eines Teilnehmers mit *Underconfidence* und einem Realismusindex von a=1,15. Die Auszahlungen werden in nach Höhe sortierter Reihenfolge dargestellt.

Wurden die Punktauszahlungen aller Teilnehmer des Experiments betrachtet, so zeigte sich der Einfluss der Korrektur besonders bei einfachen und schwierigen Items. Tabelle 9-19 zeigt die Punktsummen vor und nach der Korrektur auf der Basis des Realismusindex. Dabei wurden alle Bedingungen mit Multipler Evaluation zusammengefasst. Die einfachen Items beantworteten die Teilnehmer häufig mit einer Antwortsicherheit von 100% richtig. Da die meisten Teilnehmer *Overconfidence* zeigten, wurden diese Antwortsicherheiten und damit auch die Höhe der ausgezahlten Punkte meist reduziert. So erhielten die Teilnehmer im Durchschnitt nach der Korrektur bei einfachen Items weniger Punkte als vor der Korrektur. Die durchschnittliche Punktsumme der einfachen Items verringerte sich deshalb von 931,11 Punkten um 217,85 Punkte auf 713,26 Punkte. Die schwierigen Items konnten die Teilnehmer hingegen häufig nicht richtig lösen und gaben beispielsweise 0% Antwortsicherheit in die richtige Antwortoption wieder. In diesem Fall wurden ihnen 100 Punkte abgezogen. Durch die Korrektur wurden diese

Tabelle 9-19

Deskriptive Statistik der Punktsummen über alle Bedingungen mit Multipler Evaluation, vor und nach der Realismuskorrektur.

Schwierigkeit	Bedingung	**Mittelwert**	Standard-abweichung	Standard-fehler	Anzahl Items	N
alle	ME-korrigierte Auszahlung	**1257,38**	847,47	14,94	36	3216
	ME-nicht korrigierte Auszahlung	**1406,41**	958,73	16,91	36	3216
einfach (A)	ME-korrigierte Auszahlung	**713,26**	322,06	5,68	12	3216
	ME-nicht korrigierte Auszahlung	**931,11**	324,38	5,72	12	3216
mittel (B)	ME-korrigierte Auszahlung	**401,39**	347,67	6,13	12	3216
	ME-nicht korrigierte Auszahlung	**409,36**	452,13	7,97	12	3216
schwierig (C)	ME-korrigierte Auszahlung	**142,73**	283,82	5,00	12	3216
	ME-nicht korrigierte Auszahlung	**65,94**	388,31	6,85	12	3216

Antwortsicherheiten erhöht, und den Teilnehmern wurden in diesem Fall weniger Punkte abgezogen. Deshalb erhöhte sich die durchschnittliche Punktsumme der Teilnehmer bei schwierigen Items von durchschnittlich 65,94 Punkten um 76,79 Punkte auf 142,73 Punkte. Bei mittelschwierigen Items erzielten die Teilnehmer vor der Korrektur im Mittel 409,36 Punkte. In dieser Schwierigkeitsstufe veränderte sich die durchschnittliche Punktsumme nur geringfügig. Nach der Korrektur erhielten

die Teilnehmer durchschnittlich 7,97 Punkte weniger, nämlich 401,39 Punkte. Die Punktsumme aller Auszahlungen (36 Items) verringerte sich durch die Korrektur von 1406,41 Punkten um 149,03 Punkte auf 1257,38 Punkte. Abbildung 9-11 zeigt die Mittelwerte der korrigierten und nicht korrigierten Punktauszahlungen für alle Teilnehmer und Bedingungen des Experiments. Dargestellt werden die Mittelwerte der Punktauszahlungen je Item für alle Items des Tests und für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items. Die Tabellen mit der deskriptiven Statistik der Punktsummen der realismuskorrigierten Auszahlungen aller 36 Items und getrennt ermittelt für einfache, mittelschwierige und schwierige Items für alle Versuchsbedingungen befinden sich im Anhang A.3.

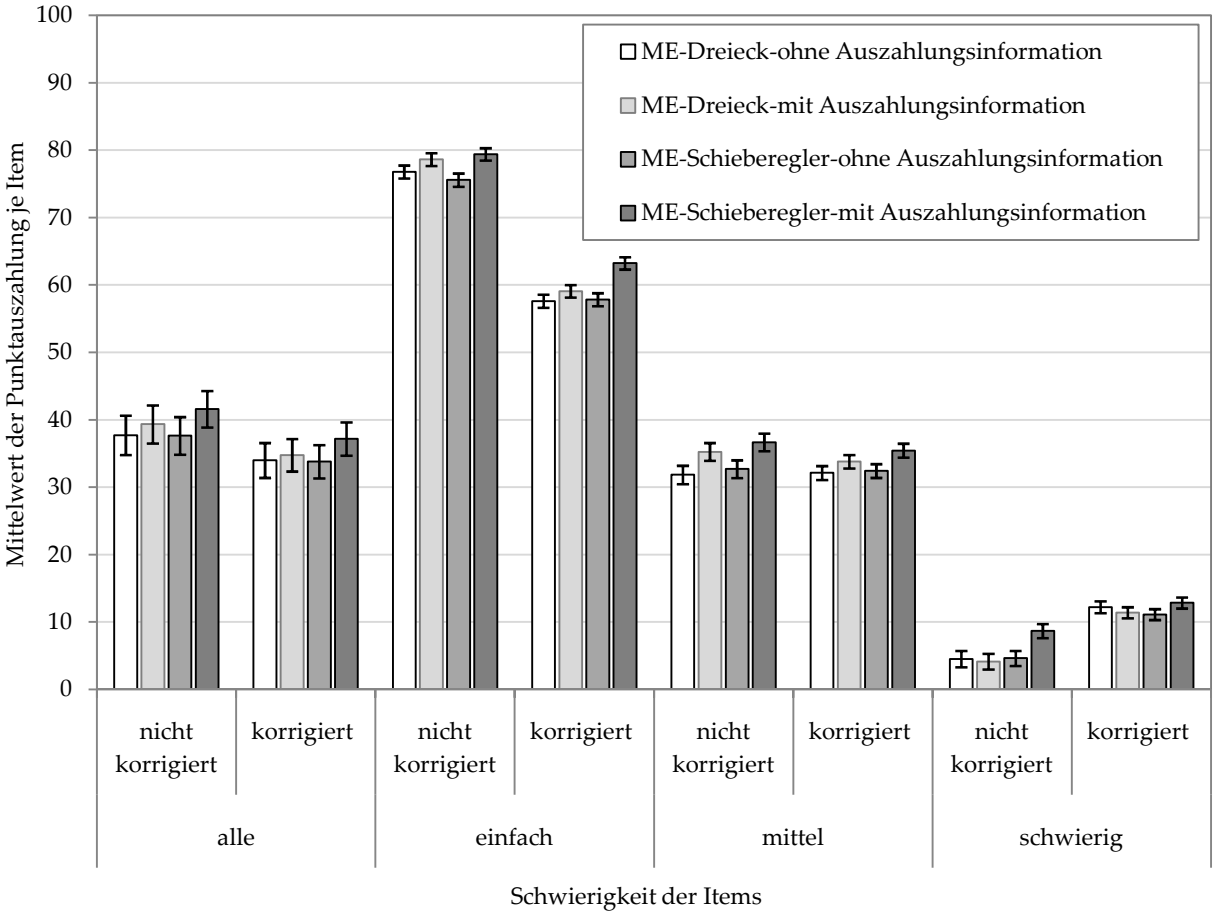

Abbildung 9-11: Mittelwerte der Punktauszahlung je Item für nicht korrigierte und korrigierte Punktauszahlungen. Die Fehlerbalken repräsentieren Standardfehler.

9.3.7 Reliabilität nach Realismuskorrektur

Auf der Basis der realismuskorrigierten Punktauszahlungen wurden die Cronbach-α-Koeffizienten zur Bestimmung der internen Konsistenz als Maß der Reliabilität berechnet. Tabelle 9-20 zeigt die α-Koeffizienten. Die Cronbach-α-Koeffizienten der korrigierten Punktauszahlungen des gesamten Tests waren in allen Bedingungen signifikant höher als die α-Koeffizienten der nicht korrigierten. Die höchste Verbesserung betrug für den gesamten Test 0,09 beispielsweise in der Bedingung „ME-Dreieck-ohne Auszahlungsinformation“, die niedrigste 0,08 in der Bedingung „ME-Schieberegler-mit Auszahlungsinformation“. Die Cronbach-α-Koeffizienten, ermittelt für nicht korrigierte und korrigierte Punkte, und die Ergebnisse der Prüfungen auf Signifikanz der Differenzen zwischen diesen Koeffizienten zeigt Tabelle 9-21.

Tabelle 9-20

Cronbach-α-Koeffizienten, berechnet für die auf der Basis des Realismusindex a korrigierten Punktauszahlungen.

Schwierigeits-stufe	Bedingung	Auszahlungsinformation	**α-korrigiert**	α für standardisierte Items	Anzahl Items
alle	ME-Dreieck	ohne	**0,93**	0,94	36
		mit	**0,92**	0,93	36
	ME-Schieberegler	ohne	**0,93**	0,93	36
		mit	**0,92**	0,93	36
einfach (A)	ME-Dreieck	ohne	**0,93**	0,94	12
		mit	**0,92**	0,93	12
	ME-Schieberegler	ohne	**0,93**	0,94	12
		mit	**0,91**	0,93	12
mittel (B)	ME-Dreieck	ohne	**0,82**	0,82	12
		mit	**0,81**	0,81	12
	ME-Schieberegler	ohne	**0,79**	0,80	12
		mit	**0,81**	0,81	12
schwierig (C)	ME-Dreieck	ohne	**0,76**	0,76	12
		mit	**0,73**	0,73	12
	ME-Schieberegler	ohne	**0,70**	0,70	12
		mit	**0,73**	0,74	12

N=804 je Bedingung.

Auch bei einfachen Items waren in allen Versuchsbedingungen die auf der Basis der korrigierten Punktauszahlungen berechneten Cronbach-α-Koeffizienten signifikant höher als die, die auf der Basis der nicht korrigierten Punktauszahlungen ermittelt wurden. Bei einfachen Items zeigten sich zudem die deutlichsten

Tabelle 9-21

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten der korrigierten und nicht korrigierten Punktauszahlungen.

Schwierigkeits-stufe	Bedingung	Auszahlungs-information	α-nicht korri-gierte Punkte	α-korri-gierte Punkte	**$\Delta\ \alpha$**	X^2	*df*	*p*
alle	ME-Dreieck	ohne	0,84	0,93	**0,09****	134,4655	1	<,0001
		mit	0,83	0,92	**0,09****	89,2447	1	<,0001
	ME-Schieberegler-	ohne	0,84	0,93	**0,09****	106,1867	1	<,0001
		mit	0,84	0,92	**0,08****	89,2447	1	<,0001
einfach (A)	ME-Dreieck	ohne	0,71	0,93	**0,22****	331,5505	1	<,0001
		mit	0,70	0,92	**0,22****	274,4532	1	<,0001
	ME-Schieberegler	ohne	0,72	0,93	**0,21****	288,5549	1	<,0001
		mit	0,73	0,91	**0,18****	201,6918	1	<,0001
mittel (B)	ME-Dreieck	ohne	0,71	0,82	**0,11****	40,9367	1	<,0001
		mit	0,67	0,81	**0,14****	51,0394	1	<,0001
	ME-Schieberegler	ohne	0,68	0,79	**0,11****	31,9439	1	<,0001
		mit	0,71	0,81	**0,10****	27,2439	1	<,0001
schwierig (C)	ME-Dreieck	ohne	0,69	0,76	**0,07****	12,2036	1	0,0008
		mit	0,63	0,73	**0,10****	17,9880	1	0,0001
	ME-Schieberegler	ohne	0,66	0,70	**0,04**	2,8022	1	0,0901
		mit	0,65	0,73	**0,08****	12,4012	1	0,0007

$\Delta\ \alpha$=Differenz zwischen α-korrigierte Punkte und α-nicht korrigierte Punkte. Positive Differenzen stellen eine Verbesserung von α für korrigierte Punkte dar, negative Differenzen eine Verschlechterung. **=signifikant auf einem Signifikanzniveau von 1%. Die Anzahl der Items betrug bei der Schwierigkeitsstufe „alle“ 36 und bei einfachen, mittelschwierigen und schwierigen Items je 12 Items. N=804 je Bedingung.

Verbesserungen von Cronbach-α durch die Korrektur, beispielsweise um 0,22 in der Bedingung „ME-Dreieck-ohne Auszahlungsinformation“. Bei mittelschwierigen Items waren die Verbesserungen der Cronbach-α-Werte insgesamt etwas geringer. Die deutlichste Verbesserung des α-Koeffizienten um 0,14 fand sich für die Bedingung „ME-Dreieck-mit Auszahlungsinformation“. Den geringsten Einfluss

zeigte die Realismuskorrektur auf die Cronbach-α-Koeffizienten der schwierigen Items. Hier betrug die höchste signifikante Verbesserung 0,10 in der Bedingung „ME-Dreieck-mit Auszahlungsinformation“. Die geringste, nicht signifikante Verbesserung von α um 0,04, zeigte die Bedingung „ME-Schieberegler-mit Auszahlungsinformation“.

Zusammenfassend wurde für alle Versuchsbedingungen mit Multipler Evaluation eine deutliche signifikante Verbesserung der Cronbach-α-Koeffizienten aufgrund der Korrektur auf der Basis des individuellen Realismusindex gezeigt.

9.3.8 Bearbeitungszeiten

Das Experiment bestand aus mehreren Abschnitten. Im Folgenden werden die Bearbeitungszeiten für diese Abschnitte getrennt betrachtet. Der erste Teil des Experiments war ein Fragebogen zur Erhebung der demographischen Daten, der zweite die Einführung des Teilnehmers in das jeweilige Antwortverfahren. In den Bedingungen mit Multipler Evaluation folgten dieser Einführung fünf Übungsaufgaben. Diesen Übungsaufgaben schloss sich der aus 36 Items bestehende Englischtest an. In den Bedingungen mit Multiple-Choice führten die Teilnehmer keine Übungen durch. Hier folgte auf die Einführung in das Antwortverfahren unmittelbar der Englischtest. Abbildung 9-12 zeigt die Bearbeitungszeiten in Minuten in diesen Abschnitten des Experiments für alle sechs Versuchsbedingungen.

Der Fragebogen zur Erhebung der demographischen Daten war in allen Versuchsbedingungen gleich gestaltet. Im Durchschnitt über die sechs Bedingungen benötigten die Testteilnehmer 2,28 Minuten (*SD*=1,85; *SE*=0,027; N=4824), um diesen Fragebogen zu bearbeiten.

Dem Fragebogen folgte die Einführung in das Antwortverfahren. Die Testteilnehmer benötigten für das Lesen dieser Einführung in den Bedingungen mit Multiple-Choice durchschnittlich 0,65 Minuten (*SD*=1,85; *SE*=0,046; N=1608). In den Bedingungen mit Multipler Evaluation war die Einführung wesentlich umfangreicher, deshalb benötigten die Teilnehmer im Durchschnitt 1,13 Minuten mehr. In den Bedingungen mit dem Dreieck betrug die Bearbeitungszeit 1,78 Minuten (*SD*=1,88; *SE*=0,047; N=1608) und in den Bedingungen mit den Schiebereglern 1,77 Minuten (*SD*=2,28; *SE*=0,057; N=1608). Wurden die Teilnehmer in

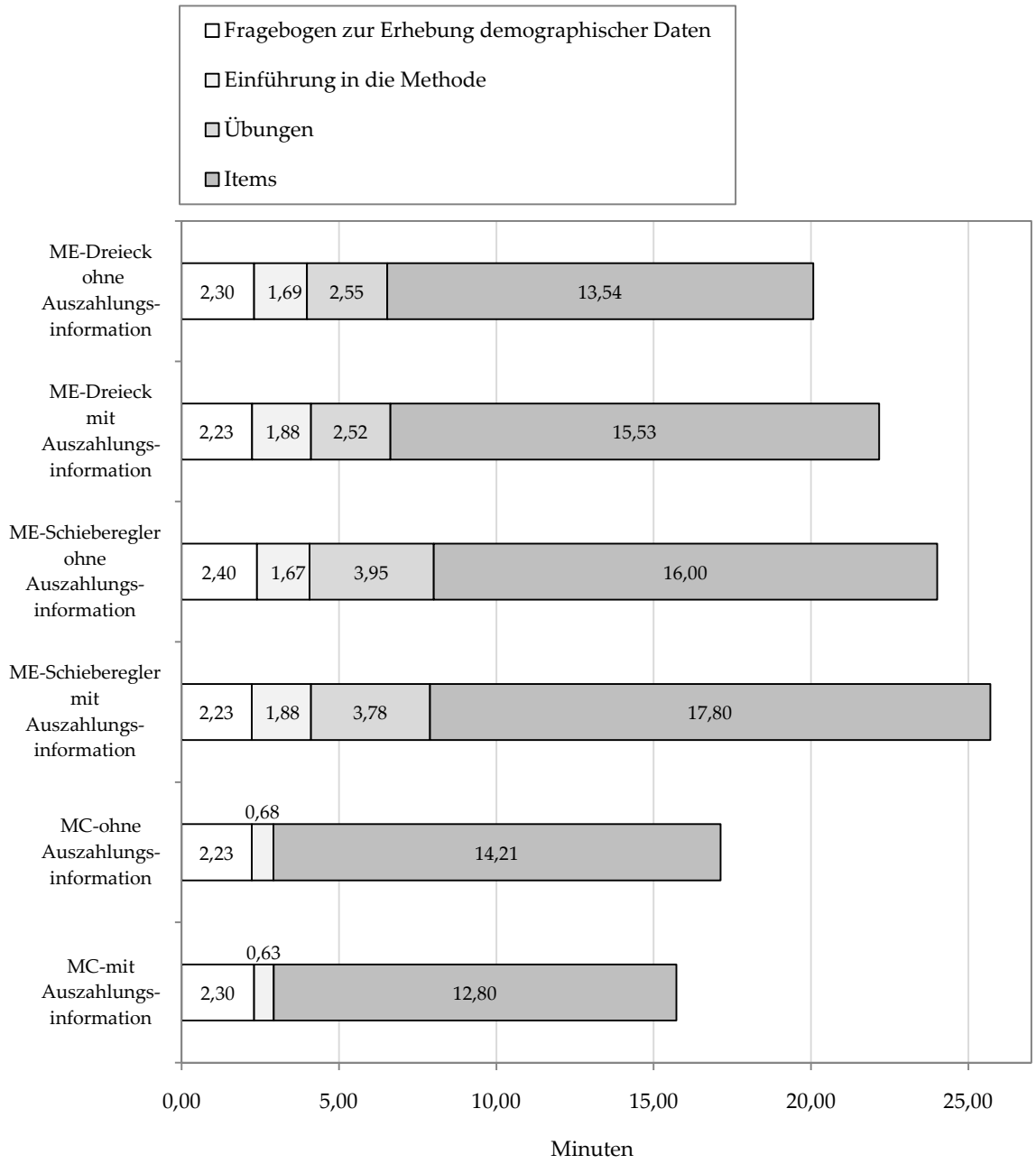

Abbildung 9-12: Durchschnittliche Bearbeitungszeiten in Minuten der vier Abschnitte des Experiments über alle sechs Bedingungen.

den Bedingungen mit Multipler Evaluation über die Auszahlungsinformation informiert, so benötigten sie durchschnittlich 1,88 Minuten (*SD*=2,33; *SE*=0,058; N=1608). Ohne eine Information über die Auszahlung dauerte das Lesen der Einführung in das Antwortverfahren im Durchschnitt 1,68 Minuten (*SD*=1,83; *SE*=0,046; N=1608), also 0,20 Minuten weniger.

Die Teilnehmer der Bedingungen mit Multipler Evaluation führten jeweils fünf Übungen durch. Für diese Übungen verwendeten sie in Abhängigkeit davon, welcher Versuchsbedingung sie zugeordnet worden waren, entweder das Dreieck oder die Schieberegler. In den Bedingungen mit dem Dreieck benötigten die Teilnehmer durchschnittlich 2,54 Minuten (*SD*=3,16; *SE*=0,079; N=1608) und mit den Schiebereglern 3,87 Minuten (*SD*=3,47; *SE*=0,087; N=1608), um die Übungen zu bearbeiten. Der Unterschied in der Bearbeitungsdauer betrug 1,33 Minuten. Eine einfaktorielle Varianzanalyse über die Bearbeitungsdauer mit dem Faktor „Antwortinstrument" zeigte, dass die durchschnittliche Bearbeitungszeit, die die Teilnehmer mit dem Dreieck für die Übungen benötigten, signifikant geringer war als die, die sie mit den Schiebereglern benötigten ($F(1, 3215)=129{,}273$; $p<{,}001$; $\eta^2={,}039$).

Die Zeitdauer, die die Testteilnehmer zur Bearbeitung des Englischtests, der aus 36 Items bestand, benötigten, variierte zwischen der kürzesten Dauer von 3,43 Minuten und der längsten von 129,68 Minuten. Tabelle 9-22 zeigt die deskriptive Statistik der Bearbeitungsdauer des Englischtests, zusammengefasst für die unabhängigen Variablen „Antwortinstrument" und „Auszahlungsinformation". Mit dem Antwortdreieck benötigten die Teilnehmer im Durchschnitt 14,53 Minuten, um alle Items zu bearbeiten. Mit den Schiebereglern dauerte diese Bearbeitung im Mittel 16,93 Minuten, also 2,40 Minuten länger. Mit den Optionsschaltflächen in den Bedingungen mit Multiple-Choice dauerte der Englischtest durchschnittlich 13,51 Minuten. Die Teilnehmer benötigten also 1,02 Minuten weniger als mit dem Dreieck

Tabelle 9-22

Deskriptive Statistik der Bearbeitungsdauer des Englischtests. Die Bedingungen wurden zusammengefasst für die unabhängigen Variablen „Antwortinstrument" und „Auszahlungsinformation".

Bedingung	Minimum	Maximum	**Mittelwert**	Standardabweichung	Standardfehler	Median	Zeit je Item	N
ME-Dreieck	3,43	129,68	**14,53**	8,79	0,22	12,35	0,40	1608
ME-Schieberegler	4,88	104,50	**16,93**	9,58	0,24	14,34	0,47	1608
MC	3,65	120,75	**13,51**	7,90	0,20	11,57	0,38	1608
ohne Auszahlungsinformation	3,43	129,68	**14,11**	9,22	0,19	11,68	0,39	2412
mit Auszahlungsinformation	3,73	89,62	**15,86**	8,48	0,17	13,64	0,44	2412

und 3,42 Minuten weniger als mit den Schiebereglern. Erhielten die Teilnehmer im Rahmen der Auszahlungsinformation ein *Feedback* nach jedem Item, so verlängerte dies die durchschnittliche Bearbeitungszeit um 1,75 Minuten von 14,11 Minuten auf 15,86 Minuten. Eine zweifaktorielle Varianzanalyse über die Bearbeitungszeiten mit dem ersten Faktor „Antwortinstrument" mit den Faktorstufen „ME-Dreieck", „ME-Schieberegler" und „MC" und über den zweiten Faktor „Auszahlungsinformation" mit den beiden Stufen „ohne Auszahlungsinformation" und „mit Auszahlungsinformation" zeigte einen signifikanten Einfluss des Antwortinstruments auf die Bearbeitungszeit der Items ($F(2, 4823)=64{,}931$; $p>0{,}001$; $\eta^2<{,}026$). Auch in Bezug auf den zweiten Faktor „Auszahlungsinformation" zeigte die Varianzanalyse einen

Tabelle 9-23

Zeitdifferenzen in Minuten und Ergebnisse der Prüfung auf Signifikanz mithilfe von Scheffé-Tests für die Bearbeitungsdauer der 36 Items des Englischtests für die unabhängige Variable „Antwortinstrument".

Bedingung 1	Mittelwert 1	Bedingung 2	Mittelwert 2	Differenz (Bedingung 2 minus Bedingung 1)	p
ME-Dreieck	14,53	ME-Schieberegler	16,93	**2,40**	<,001**
ME-Dreieck	14,53	MC	13,51	**-1,02**	,004**
ME-Schieberegler	16,93	MC	13,51	**-3,42**	<,001**

**=signifikant auf einem Signifikanzniveau von 1%, Einheit=Minuten, N je Bedingung=1608.

signifikanten Unterschied zwischen den Bedingungen ($F(1, 4823)=48{,}200$; $p>0{,}001$; $\eta^2<{,}010$). Eine signifikante Interaktion zwischen dem verwendeten Antwortinstrument und ob die Teilnehmer eine Auszahlungsinformation erhielten oder nicht bestand nicht ($F(2, 4823)=0{,}467$; $p=0{,}627$; $\eta^2<{,}001$). Tabelle 9-23 zeigt die Differenzen zwischen den Versuchsbedingungen und die Ergebnisse der Einzelvergleiche mithilfe von Scheffé-Tests. Alle Scheffé-Tests waren signifikant.

Abbildung 9-13 zeigt die Gesamtdauer des Experiments für die sechs Bedingungen und für die Bedingungen, zusammengefasst für die unabhängigen Variablen „Antwortinstrument" und „Auszahlungsinformation". Bei den Bedingungen mit Multiple-Choice ist dabei zu berücksichtigen, dass die Teilnehmer keine Übungen durchführten.

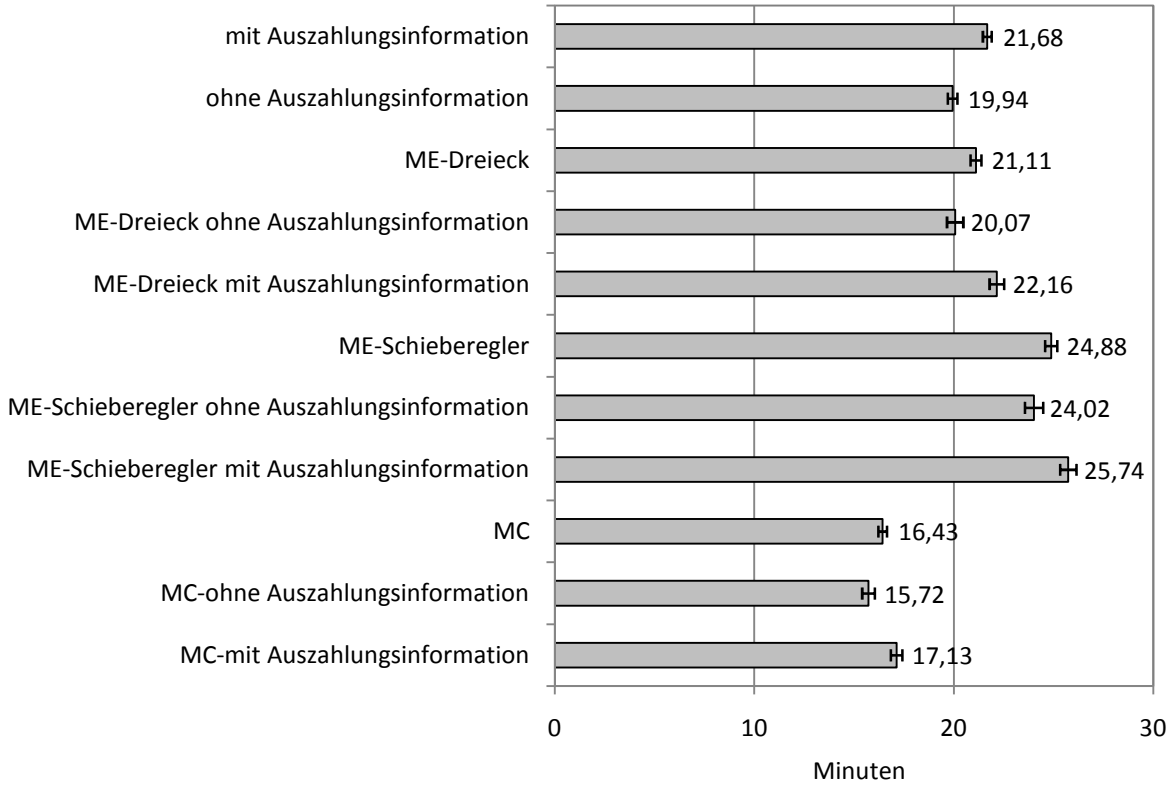

Abbildung 9-13: Durchschnittliche Gesamtdauer des Experiments in Minuten für die sechs Bedingungen und für die Bedingungen, zusammengefasst für die unabhängigen Variablen „Antwortinstrument“ und „Auszahlungsinformation“. Die Fehlerbalken repräsentieren Standardfehler.

Zusammenfassend benötigten die Teilnehmer mit den Schiebereglern eine signifikant längere Bearbeitungszeit als mit dem Dreieck. Dieses Ergebnis zeigte sich für die Übungen und für die Bearbeitung des Englischtests. Gegenüber Multiple-Choice war sowohl mit den Schiebereglern als auch mit dem Antwortdreieck die Bearbeitungszeit des Englischtests signifikant verlängert.

9.3.9 Abbruchquoten

Da die Teilnahme am Experiment anonym und ohne Kontrolle durch einen Versuchsleiter erfolgte, konnte ein Teilnehmer seine Teilnahme jederzeit beenden. Die im Folgenden als Abbruchquoten bezeichneten Prozentwerte bedeuten, dass von allen Teilnehmern (100%), die den jeweiligen Teil des Experiments aufgerufen

hatten, der entsprechende Prozentsatz von Teilnehmern nicht zum nächsten Teil des Experiments weiterging, sondern das Experiment an dieser Stelle abbrach. Der erste Teil der Testapplikation, der Fragebogen zur Erhebung der demographischen Daten, war für alle sechs Versuchsbedingungen gleich gestaltet. Während der Bearbeitung dieses Fragebogens beendeten über alle sechs Bedingungen durchschnittlich 12,31% (*SD*=0,55%) der Teilnehmer das Experiment.

Während der Einführung in das Antwortverfahren brachen in den Bedingungen mit Multipler Evaluation und dem Dreieck im Durchschnitt 3,04% (*SD*=0,75%) der Teilnehmer das Experiment ab. Während der Einführung in das Antwortverfahren Multiple Evaluation mit den Schiebereglern beendeten durchschnittlich 3,50% (*SD*=0,08) der Teilnehmer ihre Teilnahme. In den beiden Bedingungen mit Multiple-Choice betrug die Abbruchquote im Mittel 1,31% (*SD*=0,35%).

Die Teilnehmer der Bedingungen mit Multipler Evaluation führten fünf Übungsaufgaben durch. Während dieser Übungen beendeten in den Bedingungen mit dem Dreieck im Mittel 14,27% (*SD*=1,77%) der Teilnehmer das Experiment. In den Bedingungen mit den Schiebereglern brachen im Durchschnitt 20,04% (*SD*=1,88) der Teilnehmer das Experiment ab. Die Verwendung der Schieberegler für die Übungen führte also zu einer um 28,82% (Δ=5,78%) höheren Abbruchquote als die Verwendung des Dreiecks. Die Teilnehmer der Bedingung mit Multiple-Choice bearbeiteten keine Übungen.

Tabelle 9-24

Abbruchquoten für die sechs Versuchsbedingungen während der Bearbeitung der Items des Englischtests.

Bedingung	%Abbruchquote während der Bearbeitung der Items
ME-Dreieck ohne Auszahlungsinformation	7,71
ME-Dreieck mit Auszahlungsinformation	18,40
ME-Schieberegler ohne Auszahlungsinformation	9,69
ME-Schieberegler mit Auszahlungsinformation	20,67
MC-ohne Auszahlungsinformation	12,46
MC-mit Auszahlungsinformation	19,67

Im Folgenden werden die Abbruchquoten während der Bearbeitung der Items des Englischtests betrachtet. Tabelle 9-24 zeigt die Abbruchquoten in Prozent für alle Bedingungen. Tabelle 9-25 zeigt die Abbruchquoten in Prozent, zusammengefasst für die unabhängigen Variablen „Antwortinstrument“ und „Auszahlungsinformation“. In den Bedingungen, in denen die Testteilnehmer das Dreieck verwendeten, um ihre Antwortsicherheiten anzugeben, beendeten 13,05% das Experiment vorzeitig. Bei der Verwendung der Schieberegler erhöhte sich die Abbruchquote um 16,32% auf 15,18% (Δ=2,13%). Erhielten die Teilnehmer in den Bedingungen mit Multipler Evaluation keine Auszahlungsinformation, so beendeten 8,70% der Teilnehmer das

Tabelle 9-25

Abbruchquoten während der Bearbeitung der Items des Englischtests für die Bedingungen zusammengefasst für die unabhängigen Variablen „Antwortinstrument“ und „Auszahlungs-information“ und für die Bedingungen mit Multipler Evaluation und mit Multiple-Choice.

Bedingung	% Abbruchquote während der Bearbeitung der Items Mittelwert	Standard-abweichung
ME	14,12	6,37
MC	16,06	5,10
ohne Auszahlungsinformation	9,95	2,38
mit Auszahlungsinformation	19,58	1,14
ME-Dreieck	13,05	7,56
ME-Schieberegler	15,18	7,76
ME-ohne Auszahlungsinformation	8,70	1,40
ME-mit Auszahlungsinformation	19,53	1,61

Experiment während der Bearbeitung der Items. Erhielten sie eine Auszahlungsinformation, so erhöhte sich die Abbruchquote um 124,48% auf 19,53% (Δ=10,83%). In der Bedingung mit Multiple-Choice betrug die Abbruchquote 12,46%, wenn die Teilnehmer keine Auszahlungsinformation erhielten. In der Bedingung mit Multiple-Choice und einer Auszahlungsinformation betrug sie 19,67% und war damit um 57,87% höher (Δ=7,21%). Abbildung 9-14 zeigt die Abbruchquoten während der Bearbeitung der Items. Dabei wurden die Bedingungen zusammengefasst für die unabhängigen Variablen „Antwortinstrument“ und „Auszahlungsinformation“.

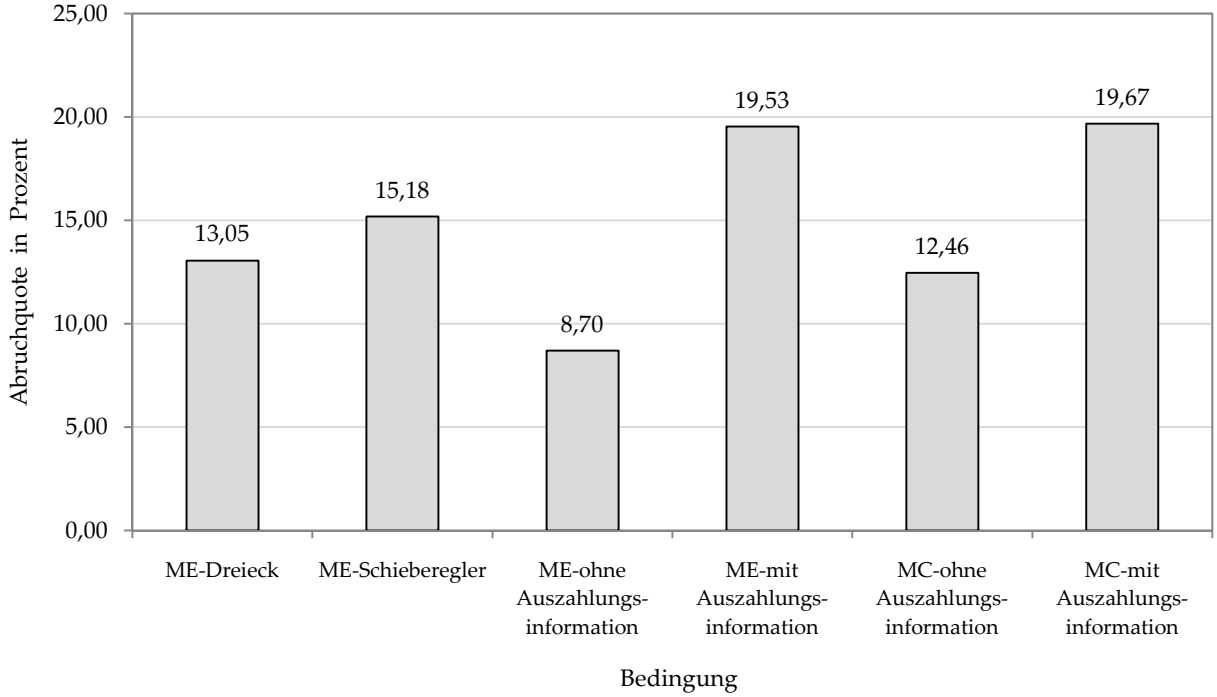

Abbildung 9-14: Abbruchquoten während der Bearbeitung der Items. Die Bedingungen wurden zusammengefasst für die unabhängigen Variablen „Antwortinstrument“ und „Auszahlungsinformation“.

Zusammenfassend beendete also eine größere Anzahl von Teilnehmern das Experiment vorzeitig, wenn sie die Schieberegler verwendeten, im Vergleich zu den Teilnehmern, die das Dreieck als Antwortinstrument nutzten. Eine Information der Teilnehmer während der Bearbeitung der Items über die Auszahlung hatte ebenfalls zur Folge, dass sich die Anzahl der Teilnehmer, die das Experiment vorzeitig beendeten, erhöhte, und zwar sowohl in den Bedingungen mit Multipler Evaluation als auch in der Bedingung mit Multiple-Choice.

9.4 Diskussion

Im ersten Experiment wurde untersucht, ob ein Englischtest, der mit dem Antwortverfahren Multiple Evaluation durchgeführt wird, eine höhere Reliabilität in Form der internen Konsistenz zeigt als ein herkömmliches Multiple-Choice-Verfahren. Die Auswertung des Tests mit Multipler Evaluation erfolgte dabei mithilfe einer logarithmischen Funktion nach Dirkzwager (2003). Der Toleranzfaktor betrug T=1 und damit der Auszahlungsbereich -100 bis 100 Punkte. Die Befunde zeigten eine signifikant höhere Reliabilität des Tests mit Multipler Evaluation im Vergleich zu dem Test mit Multiple-Choice. Um die interne Konsistenz des Tests mit Multipler Evaluation zu erreichen, müsste der Test mit Multiple-Choice um den Faktor 1,15 verlängert werden. Die Bestimmung der Verlängerungsfaktoren erfolgte nach der Methode von Spearman-Brown (Bühner, 2006). Die Ergebnisse stützen damit die Annahmen von beispielsweise Dirkzwager (2003) sowie Shuford und Brown (1975), dass der Einsatz des Antwortbewertungsverfahrens Multiple Evaluation mit einer Auswertung durch eine logarithmische Funktion zu einer Verbesserung der Reliabilität im Vergleich zu einem herkömmlichen Multiple-Choice-Test führt. Kritisch muss jedoch angemerkt werden, dass die Verbesserung der Reliabilität nur 0,02 betrug und die Signifikanzprüfung auf einer hohen Anzahl von Teilnehmern basierte (Multiple Evaluation N=3216, Multiple-Choice N=1608).

Die Varianz der beobachtbaren Testwerte aufgrund einer verfälschten Wissensreproduktion ist bei schwierigen Items in der Regel am höchsten (Abedi & Bruno, 1989). Daher war zu erwarten, dass aufgrund einer Verringerung dieser Varianz der Test mit Multipler Evaluation bei schwierigen Items die höchste Verbesserung der Reliabilität zeigen würde. Diese Annahme wurde bestätigt. Die Erhöhung der internen Konsistenz bei schwierigen Items war signifikant und entsprach einer Verlängerung des Tests mit Multiple-Choice um den Faktor 1,41. Im Gegensatz dazu war die Verbesserung bei mittelschwierigen und schwierigen Items nicht signifikant und entsprach einer Verlängerung des Tests mit Multiple-Choice um lediglich den Faktor 1,08 bzw. 1,04. Damit wurden die Befunde von Abedi und Bruno (1989) reproduziert, die eine Verbesserung der Reliabilität, besonders bei schwierigen Items, mit einer Auswertung durch eine logarithmische Funktion nach Shuford, Albert und Massengill (1966), beschrieben.

Bei einem Test mit Multipler Evaluation gibt ein Teilnehmer sein Wissen wieder, indem er alle Antwortoptionen eines Items evaluiert und seine Antwortsicherheiten berichtet. Ein Teilnehmer kann so alle Wissensstufen reproduzieren, denn er kann auch Teil- und Nichtwissen angeben. Bisher war noch nicht empirisch untersucht, wie differenziert Antwortsicherheit dabei erhoben werden muss, damit dieser Vorteil der Multiplen Evaluation im Vergleich zu Multiple-Choice sichtbar wird. Als eine weitere Fragestellung wurde deshalb untersucht, ob durch eine Erhebung mithilfe von Schiebereglern, welche die Antwortsicherheit prozentgenau erfassen können, die Reliabilität eines Tests verbessert werden kann, im Vergleich zu einer Erhebung mithilfe eines Antwortdreiecks mit nur 16 Antwortmöglichkeiten. Die Befunde zeigten keine Verbesserung der Reliabilität bei einer Erhebung von Antwortsicherheit mit den Schiebereglern im Vergleich zu einer Erhebung mit dem Dreieck. Auch wenn die Reliabilität getrennt für einfache, mittelschwierige und schwierige Items ermittelt wurde, zeigte sich kein signifikanter Einfluss des Antwortinstruments. Diese Befunde unterstützten die Vermutung von beispielsweise Leclerq (1983) sowie Paul (1993), dass eine zu differenzierte Erhebung von Antwortsicherheit nicht sinnvoll sei, da Teilnehmer nicht in der Lage seien, ihr Wissen prozentgenau anzugeben. Damit sprechen die Ergebnisse gegen die Annahme De Finettis (1965), dass durch eine differenzierte Skala eine größere Menge an Informationen gemessen werden könne. Kritisch muss bei den Schiebereglern gesehen werden, dass das genaue Einstellen prozentualer Antwortsicherheiten aufwendig und daher zeitintensiv ist, während die Teilnehmer beim Dreieck nur einmal mit der Maus in ein Feld klicken müssen, um ihre Antwortsicherheit in drei Antwortoptionen simultan anzugeben. Dies spiegelte sich in der deutlich verlängerten Bearbeitungszeit des Englischtests mit den Schiebereglern wieder. Die höhere Abbruchquote bei der Verwendung der Schieberegler im Vergleich zum Dreieck und auch zu den Bedingungen mit Multiple-Choice weist auf eine geringere Akzeptanz der Schieberegler bei den Teilnehmern hin. Es ist auch denkbar, dass manche Teilnehmer die Schieberegler nicht optimal nutzten und eher solche prozentualen Antwortsicherheiten angaben, die durch eine geringe Veränderung der Einstellung der Regler eingestellt werden konnten, anstatt ihre Antwortsicherheit prozentgenau wiederzugeben. Ein Schieberegler konnte beispielsweise auf 0% verschoben werden, um für die anderen

beiden Regler automatisch jeweils einen Wert von 50% einzustellen. Möglicherweise konnte deshalb keine Verbesserung der Reliabilität im Vergleich zum Dreieck gezeigt werden.

Im ersten Experiment wurde der Einfluss einer Information des Teilnehmers über die Punktauszahlung auf die Reliabilität eines Tests untersucht. Die Information über die Auszahlung bestand in den Experimenten aus dem *Feedforward* der zu erwartenden Punktauszahlung und dem *Feedback* der tatsächlich ausgezahlten Punkte. Brown und Shuford (1973) sowie Dirkzwager (1993) sahen beispielsweise diese Information des Teilnehmers als eine essenzielle operationale Bedingung für einen Test mit Multipler Evaluation und einer logarithmischen Auswertung an. Das Wissen über mögliche hohe Strafzahlungen soll den Teilnehmer inzentivieren, seine tatsächliche subjektive Antwortsicherheit vollkommen unverfälscht zu berichten. Die Varianz der beobachtbaren Testwerte aufgrund einer verfälschten Wiedergabe von Antwortsicherheit sollte dadurch reduziert und die Güte eines Tests verbessert werden können. Die Befunde des Experiments zeigten jedoch keinen signifikanten Einfluss der Auszahlungsinformation auf die Reliabilität des Tests. Auch bei einer getrennten Betrachtung der einfachen, mittelschwierigen und schwierigen Items zeigte sich kein Hinweis auf einen Einfluss dieser Auszahlungsinformation. Die Wirkung von *Feedback* kann durch vielfältige individuelle Faktoren beeinflusst werden (Beckmann & Beckmann, 2005; Dirkzwager, 2003; Kulhavy, Yekovich & Dyer, 1976). Die Rückmeldung von Strafzahlungen könnte daher auch unerwünschte emotionale Reaktionen wie Besorgnis oder Frustration auslösen (Bodemann, Perrez, Schär & Trepp, 2004) und so als Störgröße die Qualität eines Tests verschlechtern. Die Ergebnisse des Experiments zeigten aber auch keinen negativen Einfluss der Auszahlungsinformation auf die Reliabilität. Jedoch war die Quote der Teilnehmer, die das Experiment vorzeitig beendeten, deutlich höher, wenn sie eine Auszahlungsinformation erhielten. Es kann daher vermutet werden, dass eine Frustration der Teilnehmer, hervorgerufen durch die Rückmeldung von Strafzahlungen, der Grund für diese erhöhte Abbruchquote war. In einem Laborexperiment könnten daher die Ergebnisse konfundiert werden, da Teilnehmer in einer solchen Testsituation, beispielsweise durch die Anwesenheit eines Versuchsleiters, eher dazu animiert würden, den Test zu Ende zu führen als in einem Webexperiment (Reips, 2000a).

Für den Test mit Multiple-Choice bestätigte sich die Annahme, dass die Information über die Auszahlung keinen signifikanten Einfluss auf die Reliabilität eines Tests hat. Die Teilnehmer können diese Information nicht nutzen, denn sie haben keine andere Handlungsoption, als bei Nicht- oder Teilwissen zu raten. Damit reproduzieren die Ergebnisse auch die Befunde von Delgado und Prieto (2003) sowie Sharp, Cutler und Penrod (1988), die ebenfalls keinen positiven Einfluss von *Feedback* auf die psychometrische Qualität eines Multiple-Choice-Tests beobachteten.

Die Befunde des Experiments zeigen also, dass die Reliabilität des Tests nicht durch die Information der Teilnehmer über die Auszahlung beeinflusst wird. Es wäre daher auch möglich, Tests mit Multipler Evaluation ohne diese Information durchzuführen. Bei der Verwendung des Dreiecks als Antwortinstrument hätte dies den Vorteil, dass ein Test mit Multipler Evaluation auch als Papier-und-Bleistift-Version durchgeführt werden könnte. Bei Kalibrierungsexperimenten, die zum Ziel haben, die Kalibrierung der Teilnehmer zu verbessern, ist eine Auszahlungsinformation jedoch unbedingt erforderlich. Die Ergebnisse zeigten einen zwar geringen, aber signifikanten positiven Einfluss der Auszahlungsinformation auf den Realismusindex. Offen bleibt, ob die Teilnehmer ihre Antwortsicherheiten tatsächlich unverfälschter wiedergaben, was sie beispielsweise Lichtenstein und Fischhoff (1980) sowie Rippey und Voytovich (1982) zufolge könnten, oder, wie May (1987) annahm, ihre Antwortsicherheitsurteile nur etwas herabsetzten, um zu hohe Strafzahlungen zu vermeiden.

Noch unklar ist, wie gut kalibriert Teilnehmer ihre Antwortsicherheiten wiedergeben können. Im Experiment wurde ein individueller Realismusindex als Maß der Güte der Kalibrierung eines Teilnehmers bestimmt (Brown & Shuford, 1973; Holmes, 2002). Die meisten Teilnehmer waren nicht perfekt kalibriert, sondern zeigten *Overconfidence*. Ob die Erhebung von Antwortsicherheit mit den Schiebereglern oder mit dem Dreieck erfolgte und ob die Teilnehmer eine Information über die Auszahlung erhielten, übte einen zwar geringen, aber signifikanten Einfluss auf den Realismusindex aus. Die Teilnehmer erzielten im Durchschnitt einen höheren Realismusindex, wenn sie eine Information über die Auszahlung erhielten als ohne diese Information. Gaben die Teilnehmer ihre Antwortsicherheit durch die Schieberegler wieder, war ihr Realismusindex ebenfalls höher, als wenn sie dafür das Dreieck verwendeten. Außerdem zeigte sich eine

geringe signifikante Interaktion zwischen dem verwendeten Antwortinstrument und der Auszahlungsinformation. Den höchsten Realismusindex zeigte daher die Bedingung, in der die Teilnehmer ihre Antwortsicherheit mit den Schiebereglern wiedergaben und eine Auszahlungsinformation erhielten. Offen bleibt, was erforderlich ist, um manche Teilnehmer zu einer noch unverfälschteren Wiedergabe ihrer tatsächlichen subjektiven Antwortsicherheiten zu inzentivieren. Die Experimente beispielsweise von Lichtenstein und Fischhoff (1980) sowie von Rippey und Voytovich (1982) zeigten, dass die Kalibrierung von Teilnehmern durch Training verbessert werden kann. Im ersten Experiment dieser Arbeit führten die Teilnehmer fünf Übungen vor dem Test durch, um den Umgang mit dem Antwortinstrument zu üben und die logarithmische Auswertung kennenzulernen. Es ist denkbar, dass fünf Übungsaufgaben nicht ausreichend waren und die Teilnehmer ein längerfristiges Training benötigt hätten. Außerdem ist davon auszugehen, dass viele Testteilnehmer mit der dichotomen Antwortmethode von Multiple-Choice bereits vertraut waren und sich nicht sofort von dieser Antwortweise lösen konnten (Holmes, 2002). Kritisch muss auch gesehen werden, dass das Experiment keine echte Testsituation war, bei der das Testergebnis eine Bedeutung für den weiteren Lebensweg eines Teilnehmers haben kann. Teilnehmer könnten möglicherweise in einer realen Testsituation eher durch die Strafzahlungen zu einer unverfälschten Wiedergabe ihrer tatsächlichen subjektiven Antwortsicherheit inzentiviert werden. In einer solchen Situation wären aber auch unerwünschte Effekte durch den Abzug von Punkten, wie beispielsweise eine Verstärkung von Testangst, denkbar.

Von mehreren Autoren wurde bisher kritisiert, dass Multiple Evaluation eine deutlich längere Durchführungszeit erfordert als Multiple-Choice (Ebel, 1968; Koehler, 1971; Rippey, 1968b). Um zu prüfen, ob dieser Nachteil trotz der verbesserten technischen Möglichkeiten noch besteht, wurden im ersten Experiment die Bearbeitungszeiten der einzelnen Abschnitte des Experiments gemessen. Die Zeit, die für die Einführung in das Antwortverfahren und die Übungen benötigt wurde, war für die Bedingungen mit Multipler Evaluation deutlich länger als für die mit Multiple-Choice. Dieser Teil des Tests ist jedoch nur solange erforderlich, wie die Teilnehmer noch nicht mit dem Antwortverfahren vertraut sind. Die Teilnehmer benötigten aber auch für die Bearbeitung der Items mehr Zeit, sowohl bei der

Verwendung des Dreiecks als auch mit den Schiebereglern, gegenüber den Bedingungen mit Multiple-Choice. Die Bearbeitungszeit des Tests war dabei mit dem Dreieck deutlich kürzer als mit den Schiebereglern. Bei der Betrachtung der Bearbeitungsdauer muss aber berücksichtigt werden, dass die Teilnehmer in den Bedingungen mit Multiple-Choice sowie mit Multipler Evaluation und den Schiebereglern zum Abgeben ihrer Antwort eine zusätzliche Schaltfläche anklicken mussten. In den Bedingungen mit dem Dreieck wurden die Antwortsicherheiten hingegen schon durch Anklicken eines Feldes des Dreiecks abgegeben. Die gemessenen Zeitwerte können daher nur eine Tendenz zeigen. Ein exakter Vergleich der Bearbeitungszeiten müsste in einem weiteren Experiment erfolgen, in dem diese Unterschiede in der Implementierung der Antwortinstrumente nicht bestehen. Dies könnte beispielsweise dadurch erreicht werden, dass die Teilnehmer auch unter Verwendung des Antwortdreiecks eine Schaltfläche betätigen müssten, um zum nächsten Item zu gelangen. Dies würde jedoch für die Teilnehmer eine unnötige Erschwernis bedeuten.

Das Experiment wurde als Webexperiment durchgeführt. Da die Teilnehmer den Test anonym und ohne Kontrolle durch einen Versuchsleiter bearbeiteten, war zu erwarten, dass manche Teilnehmer ihre Teilnahme vorzeitig beenden würden (Reips, 2003). In den Bedingungen mit den Schiebereglern brachen deutlich mehr Teilnehmer das Experiment vorzeitig ab als in den Bedingungen mit dem Dreieck. Dabei beendeten die meisten Teilnehmer das Experiment bereits während der Übungen. Bei Webexperimenten können motivationale Faktoren eher zu einem vorzeitigen Abbruch der Teilnahme führen als bei Laborexperimenten (Reips, 2003). Die höhere Abbruchrate bei der Verwendung der Schieberegler ist mit einiger Wahrscheinlichkeit darauf zurückzuführen, dass Teilnehmer die Angabe ihrer Antwortsicherheit als wesentlich mühsamer empfanden als die Angabe durch das Dreieck. Sowohl in den Bedingungen mit Multipler Evaluation als auch mit Multiple-Choice beendeten zudem deutlich mehr Teilnehmer das Experiment vorzeitig, wenn sie eine Information über die Auszahlung erhielten. Die unmittelbare Leistungsrückmeldung nach jedem Item führte vermutlich zu einer Frustration von Teilnehmern und sie brachen das Experiment ab. Insgesamt zeigten die Bedingungen mit Multiple-Choice eine höhere Abbruchquote während des Englischtests als die mit Multipler Evaluation. Die Teilnehmer des Tests mit Multiple-Choice lasen nur

eine kurze Einführung und führten keine Übungen durch, ihre Aufwärmphase war also kurz (Reips, 2003). In den Bedingungen mit Multipler Evaluation mussten die Teilnehmer hingegen eine wesentlich längere Einführung in das Antwortverfahren sowie Übungen überwinden, bevor sie den Englischtest durchführen durften. Deshalb ist davon auszugehen, dass Teilnehmer, die nicht wirklich motiviert waren, das Experiment bis zum Ende durchzuführen, in den Bedingungen mit Multipler Evaluation bereits während der Einführung in das Antwortverfahren oder den Übungen ausschieden, während Teilnehmer der Bedingungen mit Multiple-Choice ihre Teilnahme erst im Englischtest beendeten. Um dies zu vermeiden, wäre es möglich gewesen, auch in den Bedingungen mit Multiple-Choice Übungen durchführen zu lassen. Dieses Vorgehen wäre aber unrealistisch für den praktischen Einsatz des Antwortverfahrens gewesen und hätte daher keinen aussagekräftigen Vergleich unter realen Einsatzbedingungen dargestellt.

Für die Daten des Experiments wurde eine nachträgliche Korrektur der Antwortsicherheiten auf der Basis des individuellen Realismusindex eines Teilnehmers vorgenommen. Durch diese Korrektur wurde eine unzureichende Kalibrierung von Teilnehmern ausgeglichen. So sollte die Varianz der beobachtbaren Testwerte, hervorgerufen durch einer verfälschten Wiedergabe der tatsächlichen subjektiven Antwortsicherheit, im Test reduziert werden. Aufgrund der Korrektur erhielt ein Teilnehmer die Antwortsicherheit, die er hätte berichten müssen, wenn er im Test perfekt kalibriert gewesen wäre. Für diese korrigierten Antwortsicherheiten wurden die Punkte erneut ausgezahlt. Die Hypothese, die auf der Basis der korrigierten Punkte ermittelte Reliabilität des Tests sei höher als die der nicht korrigierten, wurde für alle Bedingungen mit Multipler Evaluation bestätigt. Durch die Korrektur konnte beispielsweise für Multiple Evaluation mit dem Antwortdreieck und einer Information über die Auszahlung ein signifikanter Reliabilitätszugewinn erreicht werden, der einer Verlängerung des Tests mit Multipler Evaluation ohne Korrektur um den Faktor 2,53 und einer Verlängerung des Tests mit Multiple-Choice um den Faktor 2,92 entsprach.

10 Zweites Experiment

10.1 Fragestellungen und Hypothesen des zweiten Experiments

Bei einem Test mit Multipler Evaluation soll ein Teilnehmer durch die Auswertung mithilfe einer logarithmischen Funktion, die Strafzahlungen vorsieht, inzentiviert werden, sein Wissen unverfälscht zu reproduzieren. Die Höhe der Strafzahlungen kann dabei durch den Toleranzfaktor T bestimmt werden (siehe Abschnitt 4.2). Je höher der Toleranzfaktor gewählt wird, desto höher ist die Anzahl der Items, die ein Teilnehmer vollkommen richtig beantworten muss, um den Punkteverlust durch ein völlig falsch beantwortetes Item wieder aufzuwiegen. Beträgt der Toleranzfaktor beispielsweise T=3, so sind dazu drei Items erforderlich. Angesichts dessen war anzunehmen, dass ein Testteilnehmer umso mehr inzentiviert wird, seine tatsächliche subjektive Antwortsicherheit unverfälscht zu berichten, je höher der Toleranzfaktor ist, d.h. je höher die zu erwartenden Strafzahlungen sind. Erfolgt die Auswertung hingegen mithilfe einer linearen Funktion, so ist es, wie bereits ausgeführt, die beste Strategie eines Teilnehmers, immer 100% Antwortsicherheit auf die für ihn am wahrscheinlichsten richtige Antwortoption zu setzen. Zusammengefasst ist es also bei einer logarithmischen Auswertung für einen Testteilnehmer die beste Strategie, um sein Testergebnis zu maximieren, seine tatsächliche subjektive Antwortsicherheit unverfälscht zu berichten. Bei einer linearen Auswertung hingegen ist es die beste Strategie, dies nicht zu tun. Vor diesem Hintergrund war zu erwarten, dass die Varianz der beobachtbaren Testwerte eines Tests bei einer logarithmischen Auswertung im Vergleich zu einer linearen reduziert werden kann, und zwar umso mehr, je höher die Strafzahlungen sind. Als Folge daraus sollten die Reliabilität und die Validität eines Tests umso höher sein, je höher die Strafzahlungen bei einer logarithmischen Auswertung sind. Die geringste Testgüte sollte hingegen bei einer linearen Auswertung zu beobachten sein.

Denkbar wäre jedoch auch, dass Teilnehmer aufgrund einer mangelnden Kalibrierung, trotz einer Inzentivierung durch eine logarithmische Auswertung, häufig eine nur geringe Antwortsicherheit in die richtige Antwort, und somit hohe in die Distraktoren, wiedergeben. Erhielten die Testteilnehmer deshalb häufig hohe

Strafzahlungen, so könnten diese ihrerseits die Varianz der beobachtbaren Testwerte erhöhen. Dann wäre es möglich, dass bei einer Auswertung mithilfe einer linearen Funktion, die keine Strafzahlungen vorsieht, die höchste Güte eines Tests mit Multipler Evaluation beobachtet werden kann. Tatsächlich zeigt die bisherige empirische Befundlage keine klare Überlegenheit einer logarithmischen Auswertung gegenüber einer linearen. Koele, De Boo und Verschure (1987) sowie Rippey (1970) fanden beispielsweise höhere Verbesserungen der Reliabilität und der Validität des Tests bei einer Auswertung durch eine lineare im Vergleich zu einer logarithmischen Funktion. Sie führten jedoch keine Signifikanzprüfungen durch. Das Ziel des zweiten Experiments war deshalb die Klärung der Frage, ob die Reliabilität und die Validität eines Tests mit Multipler Evaluation umso mehr verbessert werden können, je höher die Strafzahlungen in der logarithmischen Auswertung sind. Im Vergleich dazu wurden der Test mit Multipler Evaluation und einer linearen Auswertung und der Test mit Multiple-Choice untersucht. Die Fragestellungen wurden mithilfe der folgenden Hypothesen geprüft:

I. Hypothese:
Je höher der Toleranzfaktor T ist, bzw. je höher die Strafzahlungen in der Auswertung bei nur geringen Antwortsicherheiten in die richtige Antwortoption sind, desto höher ist die interne Konsistenz des Tests mit Multipler Evaluation.

II. Hypothese:
Je höher der Toleranzfaktor T ist, bzw. je höher die Strafzahlungen in der Auswertung bei nur geringen Antwortsicherheiten in die richtige Antwortoption sind, desto höher ist die Validität des Tests mit Multipler Evaluation.

III. Hypothese:
Die Bedingungen mit Multipler Evaluation zeigen gegenüber den Bedingungen mit Multiple-Choice eine verbesserte interne Konsistenz.

IV. Hypothese:
Die Bedingungen mit Multipler Evaluation zeigen gegenüber den Bedingungen mit Multiple-Choice eine verbesserte Validität.

Auch im zweiten Experiment wurde untersucht, ob eine Korrektur der Antwortsicherheiten auf der Basis eines individuellen Realismusindex der Teilnehmer ein geeignetes Verfahren ist, um die Reliabilität und die Validität eines Tests nachträglich zu verbessern. Dies wurde mithilfe der folgenden Hypothesen V und VI geprüft:

V. Hypothese:
Werden die Antwortsicherheiten der Testteilnehmer auf der Basis ihres individuellen Realismusindex korrigiert, so verbessert sich die interne Konsistenz, ermittelt auf der Basis der erneuten Punktauszahlung, im Vergleich zur internen Konsistenz, ermittelt für die Punkte, die mithilfe der nicht korrigierten Antwortsicherheiten ausgezahlt wurden.

VI. Hypothese:
Werden die Antwortsicherheiten der Testteilnehmer auf der Basis ihres individuellen Realismusindex korrigiert, so verbessert sich die Validität, ermittelt auf der Basis der erneuten Punktauszahlung, im Vergleich zur Validität, ermittelt für die Punkte, die mithilfe der nicht korrigierten Antwortsicherheiten ausgezahlt wurden.

10.2 Methode

Im ersten Experiment zeigten die Teilnehmer der Bedingung „ME-Schieberegler-mit Auszahlungsinformation“ den höchsten durchschnittlichen Realismusindex. Deshalb wurden für die Erhebung von Antwortsicherheit im zweiten Experiment ausschließlich die Schieberegler verwendet. Die Teilnehmer aller Versuchsbedingungen erhielten zudem eine Information über die Auszahlung.

Für die Erhebung der Daten des zweiten Experiments wurde die Testapplikation des ersten Experiments als Vorlage verwendet. Nachfolgend werden die Änderungen in der Applikation und der Methode gegenüber dem ersten Experiment beschrieben.

10.2.1 Design

Im zweiten Experiment wurde der Englischtest mit einer Auswertung anhand einer logarithmischen Funktion (Dirkzwager, 2003) mit den Toleranzfaktoren T=0,5, T=1 und T=3 sowie einer linearen Auswertefunktion untersucht. Die Kontrollbedingung war ein herkömmliches Multiple-Choice-Verfahren. Das Versuchsdesign, in Tabelle 10-1 dargestellt, war einfaktoriell.

Tabelle 10-1

Design des 1x5-faktoriellen Versuchsplans.

Auswertefunktion
Multiple-Choice dichotom 0 oder 1 Punkt
Multiple Evaluation linear 0 bis 100 Punkte
Multiple Evaluation logarithmisch -50 bis 100 Punkte
Multiple Evaluation logarithmisch -100 bis 100 Punkte
Multiple Evaluation logarithmisch -300 bis 100 Punkte

10.2.2 Operationalisierung der unabhängigen Variable „Auswertefunktion“

In der Bedingung mit der linearen Auswertung wurde die prozentuale Antwortsicherheit eines Teilnehmers in die richtige Antwortoption als Punkte

ausgezahlt. Für die logarithmische Auswertung wurde eine Funktion nach Dirkzwager (2003) verwendet. Die Auswertung erfolgte mit den Toleranzfaktoren T=0,5, T=1 sowie T=3. Für die Umsetzung des Toleranzfaktors T in der Auswertung wird in einer logarithmischen Funktion der entsprechende Toleranzparameter t eingesetzt. Tabelle 10-2 zeigt für die Toleranzfaktoren T die entsprechenden Toleranzparameter t und den Auszahlungsbereich der jeweiligen Funktion. Abbildung 10-1 zeigt die Graphen der Auswertefunktionen.

Tabelle 10-2

Toleranzfaktor T, Toleranzparameter t und Auszahlungsbereich der vier Stufen der unabhängigen Variable „Auswertung".

Auswertefunktion	Toleranzfaktor T	Toleranzparameter t	Auszahlungsbereich
linear	-	-	0 bis 100 Punkte
logarithmisch	0,5	0,3330	-50 bis 100 Punkte
logarithmisch	1	0,1667	-100 bis 100 Punkte
logarithmisch	3	0,0134	-300 bis 100 Punkte

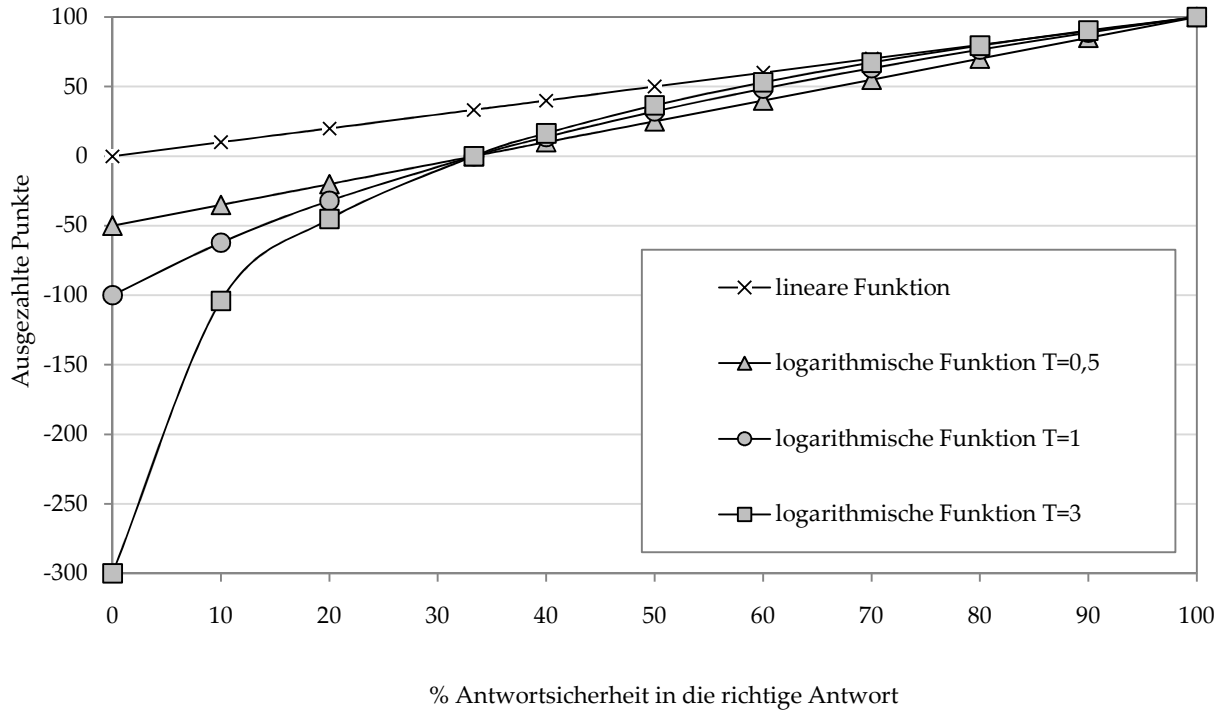

Abbildung 10-1: Graphen der vier Auswertefunktionen.

10.2.3 Operationalisierung der Selbsteinschätzung

Zur Bestimmung der externen Validität des Englischtests wurden die selbsteingeschätzten Fähigkeiten in der englischen Sprache der Teilnehmer erhoben. Die Einschätzung ihrer Fähigkeiten nahmen die Teilnehmer in den vier Kategorien „englische Sprache sprechen“, „englische Sprache verstehen“, „englische Sprache lesen“ und „englische Sprache schreiben“ vor. Abbildung 10-2 zeigt die Pole der Ratingskalen. Der niedrigste Wert jeder Skala betrug eins, der höchste sechs. Eine Ratingskala bildete die Fähigkeitsstufen des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER) ab. Der linke Pol entsprach dem Fähigkeitsniveau A1, stellte also die geringsten Fähigkeiten dar. Der rechte entsprach der Stufe C2 und damit sehr hohen Fähigkeiten. Die Teilnehmer gaben ihre selbsteingeschätzte Fähigkeit an, indem sie das entsprechende Optionsfeld markierten. Je Ratingskala konnte nur ein Feld ausgewählt werden.

Geringste Fähigkeiten (Wert 1)	**Höchste Fähigkeiten (Wert 6)**
1. Auf welchem Schwierigkeitsniveau können Sie Englisch **sprechen?**	
Sehr einfache kurze Aussagen; kann mich auf ganz einfache Art verständigen	Fließende Gespräche und Diskussionen führen, sowie Redewendungen (auch umgangs-sprachliche) verstehen
2. Auf welchem Schwierigkeitsniveau können Sie gesprochene englische Sprache **verstehen?**	
Ganz einfache Sätze, falls langsam und deutlich gesprochen	Sehr komplexe Sachverhalte, auch wenn schnell gesprochen wird
3. Auf welchem Schwierigkeitsniveau können Sie englische Texte **lesen?**	
Ganz einfache Texte	Sehr komplexe und sprachlich anspruchsvolle Texte
4. Auf welchem Schwierigkeitsniveau können Sie englische Texte **schreiben?**	
Ganz einfache Texte	Sehr komplexe und sprachlich anspruchsvolle Texte

Abbildung 10-2: Pole der Skalen zur Selbsteinschätzung der Testteilnehmer.

Die selbsteingeschätzten Fähigkeiten wurden mithilfe des Fragebogens (Teil zwei der Testapplikation) vor dem Englischtest erhoben. Die Teilnehmer konnten erst zum nächsten Teil des Experiments gelangen, wenn sie auf jeder der vier Ratingskalen ein Optionsfeld markiert hatten.

10.2.4 Beschreibung der Stichprobe

Am zweiten Experiment nahmen pro Versuchsbedingung 808 Personen teil. Die Stichprobe bestand somit aus insgesamt 4040 Testteilnehmern. Das durchschnittliche Alter in den Bedingungen betrug 33,7 Jahre (*SD*=14,0). Dabei war der jüngste Teilnehmer 10 Jahre, da Teilnehmer unter 10 Jahren aus der Auswertung ausgeschlossen wurden, und der älteste 99 Jahre. Tabelle 10-3 zeigt die deskriptive Statistik des Alters in den Versuchsbedingungen.

Tabelle 10-3

Altersverteilung in den Versuchsbedingungen.

Bedingung	Minimum	Maximum	Mittelwert	Standardfehler	Standardabweichung	Median	N
MC-0 oder 1 Punkt	12	85	**33,3**	0,498	14,2	30	808
ME-linear 0 bis 100 Punkte	12	77	**33,9**	0,497	14,1	31	808
ME-logarithmisch 50 bis 100 Punkte	13	80	**33,8**	0,497	14,1	30	808
ME-logarithmisch -100 bis 100 Punkte	13	75	**33,4**	0,490	13,9	30	808
ME-logarithmisch -300 bis 100 Punkte	11	99	**33,9**	0,484	13,7	31	808

Tabelle 10-4

Verteilung der Geschlechter in Prozent in den Versuchsbedingungen.

Versuchsbedingung	weiblich	männlich	N
MC-0 oder 1 Punkt	56,9%	43,1%	808
ME-linear 0 bis 100 Punkte	54,1%	45,9%	808
ME-logarithmisch -50 bis 100 Punkte	61,4%	38,6%	808
ME-logarithmisch -100 bis 100 Punkte	58,4%	41,6%	808
ME-logarithmisch -300 bis 100 Punkte	54,2%	45,8%	808

Die Stichprobe setzte sich aus 57,0% (*SD*=3,0%) Frauen und 43,0% (*SD*=3,0%) Männern zusammen. Tabelle 10-4 zeigt die Verteilung der Geschlechter für die Bedingungen.

10.2.5 Rekrutierung der Teilnehmer

Die Teilnehmer des zweiten Experiments wurden auf den in Tabelle 10-5 aufgeführten Webseiten auf das Experiment hingewiesen. Aufgelistet werden nur Webseiten, von denen aus mindestens 1% der Teilnehmer den Test aufriefen. Alle anderen wurden unter „Sonstige“ zusammengefasst.

Tabelle 10-5

Anteil von Teilnehmern in Prozent, die von den Webseiten (Referrer) zum Englischtest geleitet wurden.

Internetadresse	Teilnehmer in Prozent
http://www.uni-duesseldorf.de/	38,4%
http://www.google.de/ und google.com/	14,8%
Sonstige	14,4%
http://www.englisch-lernen-im-internet.de/	10,6%
http://www.testedich.de/	6,7%
http://www.lernen.sprachdirekt.de/	4,5%
http://www.eintracht.de/	4,2%
http://idw-online.de/	3,0%
http://englisch-lernen-online.net/	2,0%
http://www.zv.uni-wuerzburg.de/	1,4%

10.3 Ergebnisse

10.3.1 Schwierigkeitsstufen

Auch für das zweite Experiment wurde die Schwierigkeit der Items für die Bedingungen mit Multipler Evaluation und Multiple-Choice getrennt betrachtet. Tabelle 10-6 zeigt die deskriptive Statistik der prozentualen Antwortsicherheit in die richtige Antwort für alle Bedingungen mit Multipler Evaluation. Den richtigen Antwortoptionen der einfachen Items (Schwierigkeitsstufe A) ordneten die Testteilnehmer durchschnittlich 87,10% Antwortsicherheit je Item zu. Bei den

Tabelle 10-6

Prozentuale Antwortsicherheit in die richtige Antwort für einfache, mittelschwierige und schwierige Items für die Bedingungen mit Multipler Evaluation.

GER-Stufe	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	**Mittelwert/ Item**	Anzahl Items	N
A	einfach	279	1200	1045,24	186,06	3,27	1100,00	**87,10**	12	3232
B	mittel	100	1200	760,13	255,67	4,50	750,00	**63,34**	12	3232
C	schwierig	0	1200	543,72	217,95	3,83	482,00	**45,31**	12	3232
alle	alle	800	3600	2349,09	564,47	9,93	2325,50	**65,25**	36	3232

In jeder Schwierigkeitsstufe konnten die Teilnehmer in der Summe maximal 1200% Antwortsicherheit erreichen, insgesamt maximal 3600%.

Tabelle 10-7

Prozent richtige Antworten für einfache, mittelschwierige und schwierige Items für die Bedingungen mit Multiple-Choice.

GER-Stufe	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	**Mittelwert/ Item**	Anzahl Items	N
A	einfach	100	1200	1036,63	209,02	7,35	1100,00	**86,39**	12	808
B	mittel	0	1200	756,31	264,76	9,31	800,00	**63,03**	12	808
C	schwierig	0	1200	538,00	253,12	8,90	500,00	**45,83**	12	808
alle	alle	600	3600	2330,94	603,45	21,23	2300,00	**64,75**	36	808

In jeder Schwierigkeitsstufe konnten die Testteilnehmer in der Summe maximal 1200% richtige Antworten erreichen, insgesamt maximal 3600%.

mittelschwierigen Items (Schwierigkeitsstufe B) betrug die Antwortsicherheit in die richtige Antwort 63,34% je Item und für die schwierigen Items (Schwierigkeitsstufe C) 45,31%.

In den Bedingungen mit Multiple-Choice betrug der mittlere Schwierigkeitsindex je Item (Bühner, 2006) für einfache Items 86,39%, für mittelschwierige Items 63,03% und für schwierige Items 45,83%. Tabelle 10-7 zeigt die deskriptive Statistik der richtigen Antworten in Prozent für alle Bedingungen mit Multiple-Choice.

10.3.2 Punktauszahlungen

Als Punktsumme eines Teilnehmers wurde auch im zweiten Experiment die Summe der Punktauszahlungen aller 36 Items ermittelt. Zunächst werden die Punktsummen der Bedingungen mit Multipler Evaluation betrachtet. Eine einfaktorielle Varianzanalyse über die Punktsummen mit dem Faktor „Auswertefunktion" zeigte, dass sich die Mittelwerte der Bedingungen signifikant unterschieden ($F(3, 3228)=399{,}455$; $p<{,}001$; $\eta^2={,}271$). Je höher der Toleranzfaktor in der Auswertung war, je höher also der maximal mögliche Punktabzug, desto geringer waren die Punktsummen. Die höchste mittlere Punktsumme von 2329,77 Punkten zeigte sich deshalb bei der Auswertung durch die lineare Funktion. Die niedrigste durchschnittliche Punktsumme von 567,50 Punkten erzielten die Testteilnehmer bei der Punktauszahlung durch die Auswertefunktion mit einem Auszahlungsbereich

Tabelle 10-8

Deskriptive Statistik der Punktsummen aller 36 Items für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	**Mittelwert**	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
ME-linear 0 bis 100 Punkte	800	3600	**2329,77**	590,85	20,79	2316,00	64,72	808
ME-logarithmisch -50 bis 100 Punkte	-288	3600	**1754,62**	815,26	28,68	1699,50	48,74	808
ME-logarithmisch -100 bis 100 Punkte	-1577	3600	**1463,61**	958,92	33,73	1441,00	40,66	808
ME-logarithmisch -300 bis 100 Punkte	-5844	3600	**567,50**	1561,50	54,93	725,00	15,76	808

Die maximal erreichbare Punktsumme betrug 3600 Punkte.

von -300 bis 100 Punkten. Tabelle 10-8 zeigt die deskriptive Statistik der Punktsummen. Die höchste Differenz zwischen den Mittelwerten der Punktsummen in Höhe von 1762,27 Punkten bestand zwischen den Bedingungen „ME-linear 0 bis 100 Punkte“ und „ME-logarithmisch -300 bis 100 Punkte“. Die niedrigste Punktedifferenz in Höhe von 291,01 Punkten zeigte sich zwischen den Bedingungen „ME-logarithmisch -50 bis 100 Punkte“ und „ME-logarithmisch -100 bis 100 Punkte“. Tabelle 10-9 zeigt die Differenzen zwischen allen Versuchsbedingungen mit Multipler Evaluation und die Ergebnisse der Einzelvergleiche mithilfe von Scheffé-Tests. Die Punktsummen aller vier Bedingungen waren signifikant verschieden.

Tabelle 10-9

Punktedifferenzen und Ergebnisse der Prüfung auf Signifikanz mithilfe von Scheffé-Tests.

	ME-logarithmisch -50 bis 100 Punkte		ME-logarithmisch -100 bis 100 Punkte		ME-logarithmisch -300 bis 100 Punkte	
	Differenz	*p*	Differenz	*p*	Differenz	*p*
ME-linear 0 bis 100 Punkte	575,15**	<,001	866,16**	<,001	1762,27**	<,001
ME-logarithmisch -50 bis 100 Punkte			291,01**	<,001	1187,12**	<,001
ME-logarithmisch -100 bis 100 Punkte					896,11**	<,001

**= signifikant auf einem Niveau von 1%.

Bei einer zweifaktoriellen Varianzanalyse über die Punktsummen mit den Faktoren „Auswertefunktion“ (in vier Stufen) und „Schwierigkeit“ mit einer Messwiederholung auf dem Faktor „Schwierigkeit“ zeigte sich nach einer Greenhouse-Geisser-Korrektur ein signifikanter Einfluss der Schwierigkeit der Items auf die Punktsummen (F(1.974, 6372.314)=5073,294; p<,001; η^2=,611)). Außerdem zeigte sich, ebenfalls nach einer Greenhouse-Geisser-Korrektur, eine signifikante Interaktion zwischen der verwendeten Auswertefunktion und der Schwierigkeit der Items (F(1.974, 6372.314)=114,270; p<,001; η^2=,096). Die Mittelwerte der Punktsummen der einfachen, mittelschwierigen und schwierigen Items der Bedingungen mit Multipler Evaluation sind in Abbildung 10-3 dargestellt.

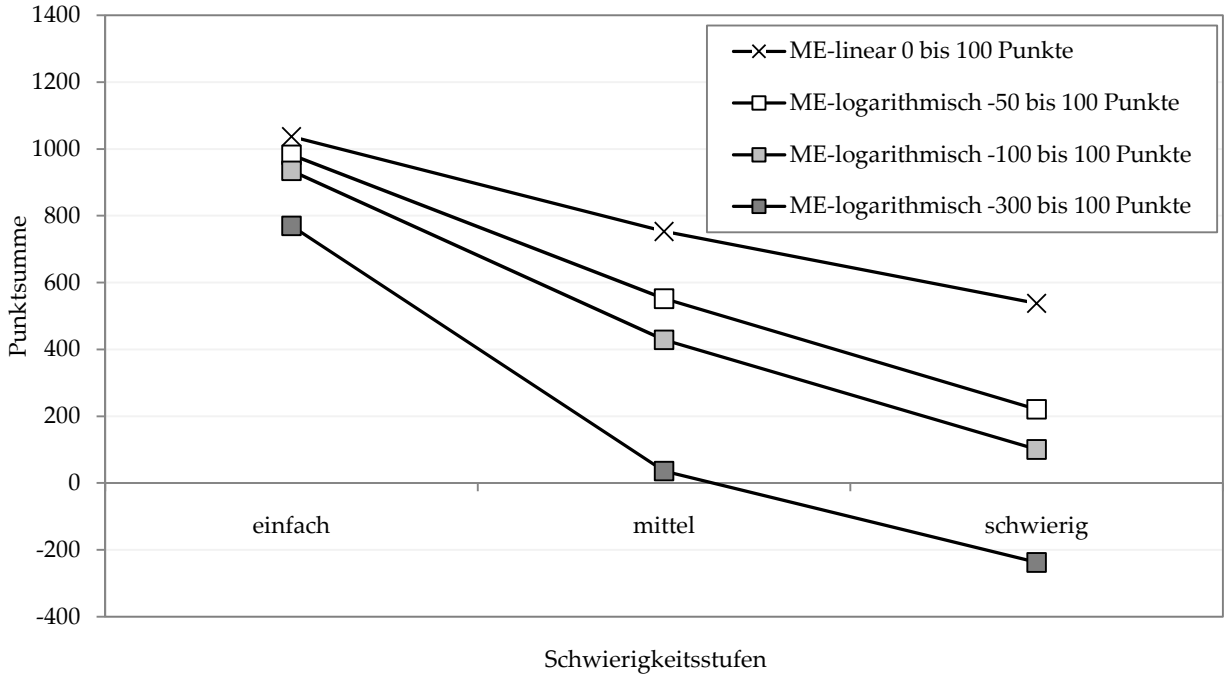

Abbildung 10-3: Mittelwerte der Punktsummen der jeweils zwölf einfachen, mittelschwierigen und schwierigen Items der Bedingungen mit Multipler Evaluation. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

In der Bedingung mit Multiple-Choice betrug der Mittelwert der Punktsumme, die die Teilnehmer für alle 36 Items des Englischtests erzielten, 23,31 Punkte. Tabelle 10-10 zeigt die deskriptive Statistik der Punktsummen der Bedingung mit Multiple-Choice.

Tabelle 10-10

Deskriptive Statistik der Punktsummen über alle 36 Items des Englischtests für die Bedingung mit Multiple-Choice.

Bedingung	Mini-mum	Maxi-mum	**Mittel-wert**	Standard-abweichung	Standard-fehler	Median	Mittelwert/ Item	N
MC-0 oder 1 Punkt	6	36	**23,31**	6,03	0,212	23,00	0,647	808

Die deskriptive Statistik der Punktsummen der einfachen, mittelschwierigen und schwierigen Items und die Mittelwerte der Punktauszahlungen je Item der Bedingung mit Multiple-Choice befinden sich in Anhang B.1.

10.3.3 Trennschärfen

Auf der Basis der Punktauszahlungen wurde für jedes Item die *part-whole-korrigierte* Trennschärfe ermittelt. Die niedrigste Trennschärfe von 0,11 zeigte sich für ein einfaches Item in der Bedingung „ME-logarithmisch -50 bis 100 Punkte“. Die höchste von 0,59 wurde für ein mittelschwieriges Item bei der Auswertung mithilfe der linearen Funktion ermittelt. Die Trennschärfen aller Items in den fünf Bedingungen befinden sich im Anhang B.2. Um die Mittelwerte der Trennschärfen bestimmen zu können, wurden diese Fisher-Z-transformiert. Die Mittelwerte wurden einer inversen Fisher-Z-Transformation unterzogen. Tabelle 10-11 zeigt die Mittelwerte und die Standardabweichungen der *part-whole-korrigierten* Trennschärfen, berechnet für alle 36 Items des Tests. Die höchste mittlere Trennschärfe von 0,40 wurde in der Bedingung mit Multipler Evaluation bei der Auswertung durch die lineare Funktion ermittelt. Je höher die mögliche Strafzahlung in den Bedingungen war, desto geringer war die durchschnittliche Trennschärfe. In der Bedingung „ME-logarithmisch -300 bis 100 Punkte“ war sie mit einem Wert von 0,28 am geringsten, und damit um 0,12 niedriger als in der Bedingung „ME-linear 0 bis 100 Punkte“. Alle

Tabelle 10-11

Mittelwert und Standardabweichung der *part-whole-korrigierten* Trennschärfen.

Bedingung	Mittelwert	Standardabweichung	Anzahl Items	N
ME-linear 0 bis 100 Punkte	0,40	0,10	36	808
ME-logarithmisch -50 bis 100 Punkte	0,38	0,10	36	808
ME-logarithmisch -100 bis 100 Punkte	0,35	0,10	36	808
ME-logarithmisch -300 bis 100 Punkte	0,28	0,08	36	808
MC-0 oder 1 Punkt	0,33	0,10	36	808

Bedingungen mit Multipler Evaluation zeigten eine höhere durchschnittliche Trennschärfe als die mit Multiple-Choice mit einem Wert von 0,33. Eine Ausnahme stellte nur die Bedingung mit der höchsten möglichen Strafzahlung von -300 Punkten

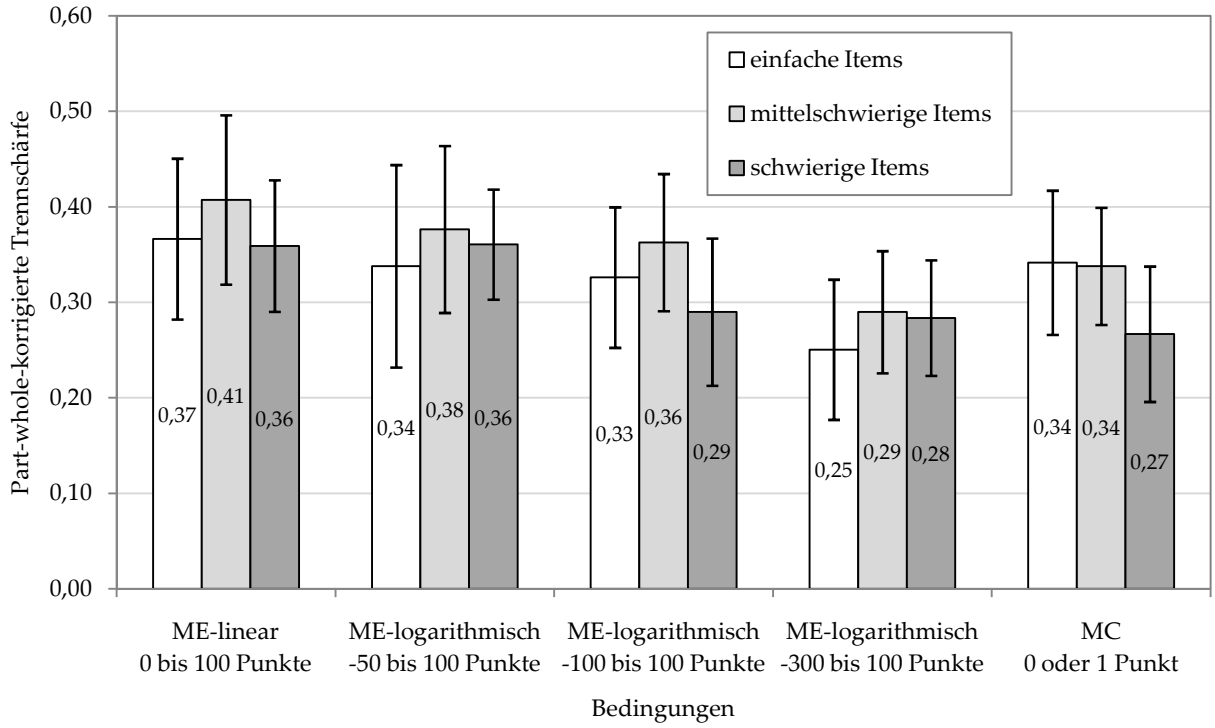

Abbildung 10-4: *Part-whole-korrigierte* Trennschärfen der einfachen, mittelschwierigen und schwierigen Items. Die Fehlerbalken repräsentieren Standardabweichungen.

dar. In dieser Bedingung war die mittlere Trennschärfe um 0,05 gegenüber der Bedingung mit Multiple-Choice verringert. Abbildung 10-4 zeigt die Trennschärfen, ermittelt für einfache, mittelschwierige und schwierige Items. Die Bedingung mit Multipler Evaluation und der Auswertung durch die lineare und die logarithmische Funktion mit einer maximalen Strafzahlung von -50 Punkten zeigte für alle Schwierigkeitsstufen eine höhere mittlere Trennschärfe als für die Bedingung mit Multiple-Choice. Wurden die Punkte durch die logarithmische Funktion mit dem Auszahlungsbereich von -100 bis 100 Punkten ausgezahlt, so war die mittlere Trennschärfe nur für einfache Items geringer als die der Bedingung mit Multiple-Choice. Für mittelschwierige und schwierige war sie hingegen verbessert. Wurden die Punkte durch die Funktion mit dem höchsten Toleranzfaktor, also einer Strafzahlung von maximal -300 Punkten ausgezahlt, so waren die Trennschärfen der einfachen und mittelschwierigen Items geringer als die der Bedingung mit Multiple-Choice. Die Trennschärfen der schwierigen Items waren dagegen in dieser Bedingung etwas höher im Vergleich zu der Bedingung mit Multiple-Choice.

10.3.4 Reliabilität

Zur Bestimmung der internen Konsistenz des Englischtests wurde der Cronbach-α-Koeffizient (Cronbach, 1951) als Maß der Reliabilität für die Punktauszahlung je Item ermittelt. Die Prüfungen auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten erfolgten mithilfe des Programms *alphatst.exe* (Lautenschlager, 1989) nach der Methode von Feldt, Woodruff und Salih (1987). Zuerst wurde die Hypothese geprüft, dass die interne Konsistenz des Tests mit Multipler Evaluation sich in Abhängigkeit von der Höhe des Toleranzfaktors in der Auswertung verbessert. Tabelle 10-12 zeigt die Cronbach-α-Koeffizienten (auch für standardisierte Items). Bei der Auswertung durch die lineare Funktion ohne Strafzahlungen war der α-Koeffizient mit α=0,89 am höchsten. Für die Versuchsbedingungen, in denen die Punkte anhand einer logarithmischen Funktion ausgezahlt wurden, verringerte sich Cronbach-α im Gegensatz zur theoretischen Erwartung umso mehr, je höher der Toleranzfaktor war. Je höher also die Strafzahlungen in der Bedingung waren, desto kleiner war Cronbach-α. Der niedrigste α-Koeffizient betrug 0,80 in der Bedingung, in der die Testteilnehmer bei 0% Antwortsicherheit in die richtige Antwortoption eine Strafzahlung von -300 Punkten erhielten. Die Bedingung mit einer maximalen Strafzahlung von

Tabelle 10-12

Cronbach-α-Koeffizienten, berechnet für alle 36 Items des Englischtests.

Bedingung	α	α für standardisierte Items	Anzahl Items	N
ME-linear 0 bis 100 Punkte	0,89	0,89	36	808
ME-logarithmisch -50 bis 100 Punkte	0,87	0,87	36	808
ME-logarithmisch -100 bis 100 Punkte	0,85	0,85	36	808
ME-logarithmisch -300 bis 100 Punkte	0,80	0,80	36	808
MC-0 oder 1 Punkt	0,83	0,84	36	808

-50 Punkten zeigte eine signifikante Verringerung der Reliabilität um 0,02 gegenüber der Bedingung mit der linearen Auswertung. Eine signifikante Verringerung gegenüber der Bedingung mit der linearen Auswertung um 0,04 bzw. 0,09 fand sich auch für die Bedingungen mit der logarithmischen Auswertung mit einer Strafzahlung von -100 bzw. -300 Punkten. In Tabelle 10-13 werden die

α-Koeffizienten nach Größe absteigend sortiert gezeigt. Die Darstellung der α-Koeffizienten in verschiedenen Spalten zeigt dabei, welche Koeffizienten sich auf einem Niveau von 5% signifikant unterschieden. Tabelle 10-14 zeigt die Einzelvergleiche der Cronbach-α-Koeffizienten für die Versuchsbedingungen.

Tabelle 10-13

Cronbach-α-Koeffizienten nach Größe absteigend sortiert. Die Darstellung der α-Koeffizienten in verschiedenen Spalten zeigt, welche Koeffizienten auf einem Niveau von 5% signifikant verschieden waren.

Bedingung	α	α	α	α
ME-linear 0 bis 100 Punkte	0,89			
ME-logarithmisch -50 bis 100 Punkte		0,87		
ME-logarithmisch -100 bis 100 Punkte			0,85	
MC-0 oder 1 Punkt			0,83	
ME-logarithmisch -300 bis 100 Punkte				0,80

Tabelle 10-14

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten, berechnet für alle 36 Items des Tests.

Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
ME-linear 0 bis 100 Punkte	0,89	ME-logarithmisch -50 bis 100 Punkte	0,87	**-0,02***	5,319	1	,0199
		ME-logarithmisch -100 bis 100 Punkte	0,85	**-0,04****	18,268	1	<,0001
		ME-logarithmisch -300 bis 100 Punkte	0,80	**-0,09****	67,094	1	<,0001
ME-logarithmisch -50 bis 100 Punkte	0,87	ME-logarithmisch -100 bis 100 Punkte	0,85	**-0,02***	3,904	1	,0453
		ME-logarithmisch -300 bis 100 Punkte	0,80	**-0,07****	35,097	1	<,0001
ME-logarithmisch -100 bis 100 Punkte	0,85	ME-logarithmisch -300 bis 100 Punkte	0,80	**-0,05***	15,726	1	,0002
MC-0 oder 1 Punkt	0,83	ME-linear 0 bis 100 Punkte	0,89	**0,06****	35,833	1	<,0001
		ME-logarithmisch -50 bis 100 Punkte	0,87	**0,04***	13,682	1	,0005
		ME-logarithmisch -100 bis 100 Punkte	0,85	**0,02**	2,986	1	,0801
		ME-logarithmisch -300 bis 100 Punkte	0,80	**-0,03***	5,032	1	,0234

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in Bedingung 2 gegenüber Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Im Vergleich zur Bedingung mit Multiple-Choice mit α=0,83 waren die Bedingungen mit Multipler Evaluation mit der linearen Auswertung ohne Strafzahlungen (α=0,89) und bei der logarithmischen Auswertung mit einer maximalen Strafzahlung von -50 Punkten (α=0,87) signifikant um 0,06 bzw. 0,04 verbessert. Erfolgte die Auswertung durch die Funktion mit einer maximalen Strafzahlung von -100 Punkten, so war α mit 0,85 nur noch um 0,02 und nicht signifikant gegenüber der Bedingung mit Multiple-Choice verbessert. Die Bedingung mit Multipler Evaluation und einer logarithmischen Auswertung mit dem Auszahlungsbereich von -300 bis 100 Punkten zeigte dagegen mit α=0,80 eine signifikante Verringerung der Reliabilität um 0,03 gegenüber der Bedingung mit Multiple-Choice.

Im Folgenden werden die Cronbach-α-Koeffizienten, ermittelt für jeweils zwölf einfache, mittelschwierige und schwierige Items, betrachtet. Auch für einfache Items zeigte sich die höchste Reliabilität mit α=0,77 in der Bedingung mit Multipler Evaluation bei der Auswertung durch die lineare Funktion. Mit steigendem Toleranzfaktor in der Auswertung verringerte sich die Reliabilität. Tabelle 10-15 zeigt die Cronbach-α-Koeffizienten (auch für standardisierte Items) für die

Tabelle 10-15

Cronbach-α-Koeffizienten der einfachen Items (Schwierigkeitsstufe A).

Bedingung	α	α für standardisierte Items	Anzahl Items	N
ME-linear 0 bis 100 Punkte	0,77	0,78	12	808
ME-logarithmisch -50 bis 100 Punkte	0,75	0,75	12	808
ME-logarithmisch -100 bis 100 Punkte	0,73	0,74	12	808
ME-logarithmisch -300 bis 100 Punkte	0,64	0,64	12	808
MC-0 oder 1 Punkt	0,75	0,76	12	808

Versuchsbedingungen. Eine signifikante Verringerung der Reliabilität um 0,04 bzw. 0,13 gegenüber der Bedingung mit der Auswertung durch die lineare Funktion zeigte sich für die Bedingungen „ME-logarithmisch -100 bis 100 Punkte“ und „ME-logarithmisch -300 bis 100 Punkte“. Tabelle 10-16 zeigt die Einzelvergleiche der Cronbach-α-Koeffizienten für die Versuchsbedingungen.

Im Vergleich zu der Bedingung mit Multiple-Choice mit einem α von 0,75 war die Reliabilität in der Bedingung mit Multipler Evaluation mit der linearen

Tabelle 10-16

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten, berechnet für einfache Items (Stufe A).

Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
ME-linear 0 bis 100 Punkte	0,77	ME-logarithmisch -50 bis 100 Punkte	0,75	**-0,02**	1,185	1	,2758
		ME-logarithmisch -100 bis 100 Punkte	0,73	**-0,04***	4,378	1	,0341
		ME-logarithmisch -300 bis 100 Punkte	0,64	**-0,13****	33,909	1	<,0001
ME-logarithmisch -50 bis 100 Punkte	0,75	ME-logarithmisch -100 bis 100 Punkte	0,73	**-0,02**	1,010	1	,3160
		ME-logarithmisch -300 bis 100 Punkte	0,64	**-0,11****	22,532	1	<,0001
ME-logarithmisch -100 bis 100 Punkte	0,73	ME-logarithmisch -300 bis 100 Punkte	0,64	**-0,09****	14,057	1	,0004
MC-0 oder 1 Punkt	0,75	ME-linear 0 bis 100 Punkte	0,77	**0,02**	1,185	1	,2758
		ME-logarithmisch -50 bis 100 Punkte	0,75	**0,00**	0	1	,3046
		ME-logarithmisch -100 bis 100 Punkte	0,73	**-0,02**	1,010	1	,3160
		ME-logarithmisch -300 bis 100 Punkte	0,64	**-0,11****	22,532	1	<,0001

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in Bedingung 2 gegenüber Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Auswertung mit α=0,77 um 0,02 verbessert, jedoch nicht signifikant. Bei der Auswertung durch die logarithmische Funktion zeigte der Test mit Multipler Evaluation bei einfachen Items keine Verbesserung der Reliabilität gegenüber Multiple-Choice. Der Cronbach-α der Bedingung „ME-logarithmisch -300 bis 100 Punkte“ war sogar signifikant um 0,11 verringert.

Bei den mittelschwierigen Items zeigte die höchste Reliabilität mit α=0,77 wiederum die Bedingung mit Multipler Evaluation bei der Auswertung durch die lineare Funktion. Tabelle 10-17 zeigt die Cronbach-α-Koeffizienten für die

Tabelle 10-17

Cronbach-α-Koeffizienten der mittelschwierigen Items (Stufe B).

Bedingung	α	α für standardisierte Items	Anzahl Items	N
ME-linear 0 bis 100 Punkte	0,77	0,77	12	808
ME-logarithmisch -50 bis 100 Punkte	0,73	0,73	12	808
ME-logarithmisch -100 bis 100 Punkte	0,72	0,72	12	808
ME-logarithmisch -300 bis 100 Punkte	0,62	0,62	12	808
MC-0 oder 1 Punkt	0,68	0,68	12	808

Versuchsbedingungen (auch für standardisierte Items). Mit steigendem Toleranzfaktor in der Auswertung verschlechterte sich die Reliabilität in allen Bedingungen signifikant. Am deutlichsten verringert, nämlich um 0,15 gegenüber der Bedingung mit der Auswertung durch die lineare Funktion, war α in der Bedingung mit der höchsten Strafzahlung von -300 Punkten. Tabelle 10-18 zeigt die Einzelvergleiche der Cronbach-α-Koeffizienten für die Versuchsbedingungen.

Im Vergleich zur Bedingung mit Multiple-Choice (α=0,68) zeigte die Bedingung mit der Auswertung durch die lineare Funktion (α=0,77) eine signifikante Verbesserung der internen Konsistenz in Höhe von 0,09. Für mittelschwierige Items hatten auch die Bedingungen mit der Auswertung durch die logarithmische Funktion bei einer maximalen Strafzahlung von -50 bzw. -100 Punkten einen signifikant höheren Wert für α. Bei der Auswertung mit der logarithmischen Funktion mit dem Auszahlungsbereich von -300 bis 100 Punkten war α um 0,06 signifikant verringert gegenüber der Bedingung mit Multiple-Choice.

Tabelle 10-18

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten, berechnet für mittelschwierige Items (Stufe B).

Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
ME-linear 0 bis 100 Punkte	0,77	ME-logarithmisch -50 bis 100 Punkte	0,73	**-0,04***	4,378	1	,0341
		ME-logarithmisch -100 bis 100 Punkte	0,72	**-0,05****	6,586	1	,0100
		ME-logarithmisch -300 bis 100 Punkte	0,62	**-0,15****	42,486	1	<,0001
ME-logarithmisch -50 bis 100 Punkte	0,73	ME-logarithmisch -100 bis 100 Punkte	0,72	**-0,01**	0,225	1	,6281
		ME-logarithmisch -300 bis 100 Punkte	0,62	**-0,13****	29,651	1	<,0001
ME-logarithmisch -100 bis 100 Punkte	0,72	ME-logarithmisch -300 bis 100 Punkte	0,62	**-0,10****	15,848	1	,0002
MC-0 oder 1 Punkt	0,68	ME-linear 0 bis 100 Punkte	0,77	**0,09****	18,501	1	,0001
		ME-logarithmisch -50 bis 100 Punkte	0,73	**0,05***	4,920	1	,0249
		ME-logarithmisch -100 bis 100 Punkte	0,72	**0,04**	3,041	1	,0773
		ME-logarithmisch -300 bis 100 Punkte	0,62	**-0,06***	5,034	1	,0233

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in Bedingung 2 gegenüber Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Tabelle 10-19 zeigt die Cronbach-α-Koeffizienten der schwierigen Items für alle Versuchsbedingungen. Für schwierige Items war die Reliabilität in den Bedingungen „ME-linear 0 bis 100 Punkte“ und „ME-logarithmisch -50 bis 100 Punkte“ jeweils mit α=0,75 am höchsten. Die niedrigste Reliabilität hatte mit α=0,60 die Bedingung mit Multiple-Choice. Die Bedingungen „ME-logarithmisch -100 bis 100 Punkte“ und „ME-logarithmisch -300 bis 100 Punkte“ hatten gegenüber der Bedingung mit der

Tabelle 10-19

Cronbach-α-Koeffizienten der schwierigen Items (Stufe C).

Bedingung	α	α für standardisierte Items	Anzahl Items	N
ME-linear 0 bis 100 Punkte	0,75	0,75	12	808
ME-logarithmisch -50 bis 100 Punkte	0,75	0,75	12	808
ME-logarithmisch -100 bis 100 Punkte	0,66	0,66	12	808
ME-logarithmisch -300 bis 100 Punkte	0,67	0,68	12	808
MC-0 oder 1 Punkt	0,60	0,60	12	808

Tabelle 10-20

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten, berechnet für schwierige Items (Stufe C).

Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
ME-linear 0 bis 100 Punkte	0,75	ME-logarithmisch -50 bis 100 Punkte	0,75	**0,00**	0	1	,3046
		ME-logarithmisch -100 bis 100 Punkte	0,66	**-0,09****	16,066	1	,0002
		ME-logarithmisch -300 bis 100 Punkte	0,67	**-0,08***	13,109	1	,0006
ME-logarithmisch -50 bis 100 Punkte	0,75	ME-logarithmisch -100 bis 100 Punkte	0,66	**-0,09****	16,050	1	,0002
		ME-logarithmisch -300 bis 100 Punkte	0,67	**-0,08***	13,095	1	,0006
ME-logarithmisch -100 bis 100 Punkte	0,66	ME-logarithmisch -300 bis 100 Punkte	0,67	**-0,01**	0,152	1	,6675
MC-0 oder 1 Punkt	0,60	ME-linear 0 bis 100 Punkte	0,75	**0,15****	37,283	1	<,0001
		ME-logarithmisch -50 bis 100 Punkte	0,75	**0,15****	37,283	1	<,0001
		ME-logarithmisch -100 bis 100 Punkte	0,66	**0,06***	4,498	1	,0318
		ME-logarithmisch -300 bis 100 Punkte	0,67	**0,07***	6,299	1	,0117

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Auswertung durch die lineare Funktion eine um 0,09 bzw. 0,08 signifikant geringere interne Konsistenz. Tabelle 10-20 zeigt die Einzelvergleiche der Cronbach-α-Koeffizienten für die Versuchsbedingungen.

Die Reliabilität der schwierigen Items war in allen Versuchsbedingungen mit Multipler Evaluation signifikant im Vergleich zur Reliabilität der Bedingung mit Multiple-Choice verbessert.

Zusammenfassend wurde die Hypothese, dass die Bedingungen mit Multipler Evaluation eine Verbesserung der Reliabilität zeigen, je höher der Toleranzfaktor T in der Auswertung ist, nicht gestützt. Stattdessen verschlechterte sich die Reliabilität umso mehr, je höher die Strafzahlungen waren. Die Reliabilität des Tests mit Multipler Evaluation zeigte hingegen die deutlichste signifikante Verbesserung im Vergleich zu Multiple-Choice, wenn die Punkte durch die lineare Funktion ausgezahlt wurden. Nur wenn ausschließlich die schwierigen Items betrachtet wurden, zeigten alle Bedingungen mit Multipler Evaluation eine höhere Reliabilität als die mit Multiple-Choice.

10.3.5 Validität

Vor der Bearbeitung des Englischtests schätzten die Teilnehmer ihre Fähigkeiten in der englischen Sprache auf einer Skala von eins bis sechs in den vier Kategorien „englische Sprache sprechen“, „englische Sprache verstehen“, „englische Sprache lesen“ und „englische Sprache schreiben“ ein. Um zu prüfen, ob signifikante Unterschiede in Bezug auf die selbsteingeschätzte Fähigkeit der Teilnehmer in den vier Kategorien bestanden, wurde pro Kategorie eine einfaktorielle Varianzanalyse über die selbsteingeschätzte Fähigkeit aller Bedingungen des Experiments mit dem Faktor „Auswertung“ berechnet.

Am höchsten, mit durchschnittlich 3,91 (*SD*=1,27; *SE*=0,020; N=4040), schätzten die Testteilnehmer ihre Fähigkeit ein, die englische Sprache lesen zu können. Eine einfaktorielle Varianzanalyse über die selbsteingeschätzte Fähigkeit „englische Sprache lesen“ zeigte keine signifikanten Unterschiede zwischen den Bedingungen ($F(4, 4039)=,608$; $p=,657$; $\eta^2=,001$). Ihre Fähigkeiten, die englische Sprache zu verstehen, beurteilten alle Teilnehmer im Durchschnitt mit 3,68 (*SD*=1,23; *SE*=0,091; N=4040). Auch für diese Selbsteinschätzung zeigte eine einfaktorielle Varianzanalyse

keinen signifikanten Unterschied zwischen den Bedingungen (F(4, 4039)=,606; p=,658; η^2=,001). Der Mittelwert der selbsteingeschätzten Fähigkeit aller Testteilnehmer in der Kategorie „englische Sprache sprechen" betrug 3,28 (SD=1,31; SE=0,021; N=4040). Eine einfaktorielle Varianzanalyse über die selbsteingeschätzte Fähigkeit „englische Sprache sprechen" zeigte keine signifikanten Unterschiede zwischen den Versuchsbedingungen (F(4, 4039)=0,912; p=,456; η^2=,001). Am geringsten, nämlich im Durchschnitt mit 3,05 (SD=1,20; SE=0,019; N=4040), schätzten die Testteilnehmer ihre Fähigkeit ein, die englische Sprache schreiben zu können. Ein signifikanter Unterschied zwischen den Bedingungen zeigte sich in einer einfaktoriellen Varianzanalyse für die selbsteingeschätzte Fähigkeit „englische Sprache schreiben" ebenfalls nicht (F(4, 4039)=0,918; p=,452; η^2=,001).

Als Maß der externen Validität wurde die Korrelation (nach Pearson) der selbsteingeschätzten Fähigkeit mit den Punktsummen der Testteilnehmer ermittelt. Dazu wurden die Daten z-transformiert (Bortz, 2005, S. 45) und der Mittelwert über alle vier Kategorien der Selbsteinschätzung gebildet. Für diesen Mittelwert (z-Index) wurde die Korrelation r mit den Punktsummen der Teilnehmer bestimmt. Für die Prüfung auf signifikante Unterschiede der Korrelationskoeffizienten wurde die Methode nach Bortz (2005, S. 220 f.) angewendet. Dazu wurden die Koeffizienten Fisher-Z-transformiert. Tabelle 10-21 zeigt die Korrelationen und die Z-transformierten Korrelationskoeffizienten des z-Index der Selbsteinschätzung mit den Punktsummen, ermittelt für alle 36 Items des Tests. In Tabelle 10-22 werden die

Tabelle 10-21

Korrelationskoeffizienten und Fisher-Z-transformierte Koeffizienten des z-Index der Selbsteinschätzung mit den Punktsummen.

Bedingung	r	Z(r)	N
ME-linear 0 bis 100 Punkte	,67	0,81	808
ME-logarithmisch -50 bis 100 Punkte	,65	0,78	808
ME-logarithmisch -100 bis 100 Punkte	,60	0,70	808
ME-logarithmisch -300 bis 100 Punkte	,50	0,54	808
MC-0 oder 1 Punkt	,60	0,69	808

Alle Korrelationen waren auf einem Niveau von p<,001 signifikant, Anzahl der Items=36.

Korrelationskoeffizienten nach Größe absteigend sortiert gezeigt. Die Darstellung in verschiedenen Spalten zeigt dabei, welche Koeffizienten auf einem Niveau von 5%

signifikant verschieden waren. Die Korrelation der selbsteingeschätzten Fähigkeit mit den Punktsummen zeigte je nach Versuchsbedingung eine mittlere bis hohe Validität (Fisseni, 1997). Hoch war die Validität mit r=,67, wenn die Punkte anhand der linearen Funktion ohne Strafzahlung ausgezahlt wurden. Je höher der Toleranzfaktor in der Auswertung war, je höher also die Strafzahlungen waren,

Tabelle 10-22

Fisher-Z-transformierte Korrelationskoeffizienten und Korrelationskoeffizient, nach Größe absteigend sortiert. Die Darstellung der Koeffizienten in verschiedenen Spalten zeigt, welche Koeffizienten auf einem Niveau von 5% signifikant voneinander verschieden waren.

Bedingung	Z(r)	r	Z(r)	r	Z(r)	r
ME-linear 0 bis 100 Punkte	0,81	,67				
ME-logarithmisch -50 bis 100 Punkte	0,78	,65				
ME-logarithmisch -100 bis 100 Punkte			0,70	,60		
MC-0 oder 1 Punkt			0,69	,60		
ME-logarithmisch -300 bis 100 Punkte					0,54	,50

desto niedriger war die Validität. Mit der Auswertung durch die logarithmische Funktion mit dem Auszahlungsbereich von -300 bis 100 Punkten betrug r=,50 und zeigte damit eine mittlere Validität. Auch die Bedingung mit Multiple-Choice zeigte mit r=,60 eine mittlere Validität. Für den Vergleich der Korrelationskoeffizienten zwischen den Bedingungen wurden die Z-transformierten Koeffizienten betrachtet. Die Bedingung mit Multipler Evaluation und der Auszahlung der Punkte anhand der linearen Funktion ohne Strafzahlungen zeigte mit Z(r)=0,81 (r=,67) den höchsten Korrelationskoeffizienten. Bei der Auszahlung der Punkte durch die logarithmische Funktion mit einem geringen Toleranzfaktor von T=0,5 (Auszahlungsbereich -50 bis 100 Punkte) verringerte sich demgegenüber die Korrelation nicht signifikant um 0,03 auf Z(r)=0,78 (r=,65). Tabelle 10-23 zeigt die Ergebnisse der Einzelprüfungen auf signifikante Unterschiede zwischen den Korrelationskoeffizienten. Eine signifikante Verringerung der Validität um 0,11 gegenüber der Bedingung mit der linearen Auswertung fand sich in der Bedingung „ME-logarithmisch -100 bis 100 Punkte“ mit einem Z(r) von 0,70 (r=,60). Eine signifikante Verringerung der Korrelation um 0,24, im Vergleich zur Bedingung mit der linearen Auswertung, zeigte die Bedingung mit

Tabelle 10-23

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen (nach Bortz, 2005) für die Daten der 36 Items.

Bedingung 1	Z(r)1	Bedingung 2	Z(r)2	**Δ Z(r)**	z	*df*	*p*
ME-linear 0 bis 100 Punkte	0,81	ME-logarithmisch -50 bis 100 Punkte	0,78	**-0,03**	-0,602	1	n.s.
		ME-logarithmisch -100 bis 100 Punkte	0,70	**-0,11***	-2,207	1	,05
		ME-logarithmisch -300 bis 100 Punkte	0,54	**-0,27****	-5,417	1	<,0005
ME-logarithmisch -50 bis 100 Punkte	0,78	ME-logarithmisch -100 bis 100 Punkte	0,70	**-0,08**	-1,605	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,54	**-0,24****	-4,815	1	<,0005
ME-logarithmisch -100 bis 100 Punkte	0,70	ME-logarithmisch -300 bis 100 Punkte	0,54	**-0,16***	-3,210	1	,0,005
MC-0 oder 1 Punkt	0,69	ME-linear 0 bis 100 Punkte	0,81	**0,12****	2,407	1	,01
		ME-logarithmisch -50 bis 100 Punkte	0,78	**0,09***	1,806	1	,05
		ME-logarithmisch -100 bis 100 Punkte	0,70	**0,01**	0,201	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,54	**-0,15****	-3,009	1	,005

Z(r)1=Fisher-Z-transformierter Korrelationskoeffizient der ersten Bedingung des Einzelvergleichs, Z(r)2=Fisher-Z-transformierter Korrelationskoeffizient der zweiten Bedingung, Δ Z(r)=Differenz zwischen Bedingung 1 und 2 (Z(r)2-Z(r)1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in Bedingung 2 gegenüber Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

der Auswertung mithilfe des höchsten Toleranzfaktors (T=3) mit einem Z(r) von 0,54 (r=,50). Die Validität der Bedingung „ME-logarithmisch -50 bis 100 Punkte“ war gegenüber den Bedingungen „ME-logarithmisch -100 bis 100 Punkte“ und „ME-logarithmisch -300 bis 100 Punkte“ verbessert, signifikant jedoch nur für die Bedingung „ME-logarithmisch -300 bis 100 Punkte“. Die Bedingung „ME-logarithmisch -100 bis 100 Punkte“ zeigte eine um 0,16 signifikant höhere Validität als die Bedingung „ME-logarithmisch -300 bis 100 Punkte“.

Der Vergleich zwischen Multiple-Choice und Multipler Evaluation zeigte für die Bedingung mit der linearen Auswertung und einem Z(r) von 0,81 (r=,67) eine signifikante Verbesserung um 0,12 gegenüber der Bedingung mit Multiple-Choice und einem Z(r) von 0,69 (r=,60). Auch mit der Auswertung mithilfe der logarithmischen Funktion und einer Strafzahlung von maximal -50 Punkten war die Validität mit Z(r)=0,78 (r=,65) gegenüber der Bedingung mit Multiple-Choice signifikant verbessert. Eine um 0,01 nur geringfügig verbesserte Validität gegenüber der Validität der Bedingung mit Multiple-Choice zeigte mit Z(r)=0,70 (r=,60) die Bedingung mit einer Strafzahlung von maximal -100 Punkten. Um 0,15 signifikant

verschlechtert war der Korrelationskoeffizient mit Z(r)=0,54 (r=,50) in der Bedingung mit der Auswertung durch die Funktion mit dem Auszahlungsbereich von -300 bis 100 Punkten.

Im Folgenden werden die Korrelationen für die jeweils zwölf Items der einfachen, mittelschwierigen und schwierigen Items betrachtet. Tabelle 10-24 zeigt die Korrelationskoeffizienten und die Fisher-Z-transformierten Koeffizienten. Auch für einfache Items zeigte sich mit r=,58 (Z(r)=0,66) die höchste Validität für die Bedingungen mit Multipler Evaluation bei der Auswertung durch die lineare Funktion, die keine Strafzahlungen vorsieht. Alle Bedingungen, in denen die Auswertung der Punkte anhand der logarithmischen Funktion erfolgte, zeigten geringere Korrelationskoeffizienten. Die deutlichste signifikante Verschlechterung des Korrelationskoeffizienten um 0,13, im Vergleich zum Koeffizienten der Bedingung mit der linearen Auswertung (Z(r)=0,66, r=,68), zeigte die Bedingung „ME-logarithmisch -300 bis 100 Punkte“ (Z(r)=0,53, r=,49). Tabelle 10-25 können die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Fisher-Z-transformierten Korrelationskoeffizienten entnommen werden.

Tabelle 10-24

Korrelation des z-Index der Selbsteinschätzung mit den Punktsummen und Fisher-Z-transformierte Korrelationskoeffizienten der jeweils zwölf einfachen, mittelschwierigen und schwierigen Items.

Schwierigkeit	einfach (A)		mittel (B)		schwierig (C)		
Bedingung	r	Z(r)	r	Z(r)	r	Z(r)	N
ME-linear 0 bis 100 Punkte	,58	0,66	,62	0,73	,52	0,57	808
ME-logarithmisch -50 bis 100 Punkte	,56	0,64	,58	0,66	,53	0,59	808
ME-logarithmisch -100 bis 100 Punkte	,51	0,56	,56	0,63	,43	0,46	808
ME-logarithmisch -300 bis 100 Punkte	,49	0,53	,44	0,48	,27	0,28	808
MC-0 oder 1 Punkt	,53	0,59	,53	0,59	,44	0,47	808

Alle Korrelationen waren auf einem Niveau von p<,001 signifikant.

Beim Vergleich der Bedingungen mit Multipler Evaluation mit der Bedingung mit Multiple-Choice zeigte sich für einfache Items eine Verbesserung der Validität für die Bedingungen „ME-linear 0 bis 100 Punkte“ und „ME-logarithmisch -50 bis 100 Punkte“ jeweils um 0,05. Jedoch waren diese nicht signifikant. Eine ebenfalls nicht signifikante Verringerung des Wertes für Z(r) zeigten die Bedingungen mit der

Auswertung durch die logarithmische Funktion und einer Strafzahlung von maximal -100 Punkten sowie maximal -300 Punkten.

Tabelle 10-25

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen (nach Bortz, 2005) für die Daten der zwölf einfachen Items (Stufe A).

Bedingung 1	Z(r)1	Bedingung 2	Z(r)2	**ΔZ(r)**	z	*df*	*p*
ME-linear 0 bis 100 Punkte	0,66	ME-logarithmisch -50 bis 100 Punkte	0,64	**-0,02**	-0,401	1	n.s.
		ME-logarithmisch -100 bis 100 Punkte	0,56	**-0,10***	-2,006	1	,05
		ME-logarithmisch -300 bis 100 Punkte	0,53	**-0,13****	-2,608	1	,005
ME-logarithmisch -50 bis 100 Punkte	0,64	ME-logarithmisch -100 bis 100 Punkte	0,56	**-0,08**	-1,605	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,53	**-0,11***	-2,207	1	,05
ME-logarithmisch -100 bis 100 Punkte	0,56	ME-logarithmisch -300 bis 100 Punkte	0,53	**-0,03**	-0,602	1	n.s.
MC-0 oder 1 Punkt	0,59	ME-linear 0 bis 100 Punkte	0,66	**0,07**	1,404	1	n.s.
		ME-logarithmisch -50 bis 100 Punkte	0,64	**0,05**	1,003	1	n.s.
		ME-logarithmisch -100 bis 100 Punkte	0,56	**-0,03**	-0,602	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,53	**-0,06**	-1,204	1	n.s.

Z(r)1=Fisher-Z-transformierter Korrelationskoeffizient der ersten Bedingung des Einzelvergleichs, Z(r)2=Fisher-Z-transformierter Korrelationskoeffizient der zweiten Bedingung, Δ Z(r)=Differenz zwischen Bedingung eins und zwei (Z(r)2- Z(r)1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in Bedingung zwei gegenüber der Bedingung eins, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von1%.

Die Daten der mittelschwierigen Items zeigten ebenfalls die höchste Validität Z(r)=0,73 (r=,62) für die Bedingung mit Multipler Evaluation und der Auswertung durch die lineare Funktion ohne Strafzahlungen. Die Ergebnisse der Prüfung auf signifikante Unterschiede zeigt Tabelle 10-26. Die Bedingungen mit einer logarithmischen Auswertung zeigten auch bei mittelschwierigen Items eine Verringerung der Validität im Vergleich zu der Bedingung mit der linearen Auswertung, und zwar in Abhängigkeit von der Höhe der möglichen Strafzahlung. Die niedrigste Validität und damit eine signifikante Verringerung um 0,25 gegenüber der linearen Auswertung zeigte deshalb die Bedingung „ME-logarithmisch -300 bis 100 Punkte“ mit einem Z(r) von 0,48 (r=,44).

Im Vergleich zu Multiple-Choice (Z(r)=0,59; r=,53) zeigte sich für mittelschwierige Items eine signifikante Verbesserung der Validität um 0,14 für die Bedingung mit Multipler Evaluation und der Auswertung durch die lineare

Tabelle 10-26

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen (nach Bortz, 2005) für die Daten der zwölf mittelschwierigen Items (Stufe B).

Bedingung 1	Z(r)1	Bedingung 2	Z(r)2	ΔZ(r)	z	*df*	*p*
ME-linear 0 bis 100 Punkte	0,73	ME-logarithmisch -50 bis 100 Punkte	0,66	-0,07	-1,404	1	n.s.
		ME-logarithmisch -100 bis 100 Punkte	0,63	-0,10*	-2,006	1	,05
		ME-logarithmisch -300 bis 100 Punkte	0,48	-0,25**	-5,016	1	<,0005
ME-logarithmisch -50 bis 100 Punkte	0,66	ME-logarithmisch -100 bis 100 Punkte	0,63	-0,03	-0,602	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,48	-0,18**	-3,611	1	<,0005
ME-logarithmisch -100 bis 100 Punkte	0,63	ME-logarithmisch -300 bis 100 Punkte	0,48	-0,15**	-3,009	1	,005
MC-0 oder 1 Punkt	0,59	ME-linear 0 bis 100 Punkte	0,73	0,14**	2,809	1	,005
		ME-logarithmisch -50 bis 100 Punkte	0,66	0,07	1,404	1	n.s.
		ME-logarithmisch -100 bis 100 Punkte	0,63	0,04	0,802	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,48	-0,11*	-2,207	1	,05

Z(r)1=Fisher-Z-transformierter Korrelationskoeffizient der ersten Bedingung des Einzelvergleichs, Z(r)2=Fisher-Z-transformierter Korrelationskoeffizient der zweiten Bedingung, Δ Z(r)=Differenz zwischen der Bedingung eins und zwei (Z(r)2-Zr(1)). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in Bedingung zwei gegenüber der Bedingung eins, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Funktion (Z(r)=0,73; r=,62). Auch die Korrelationskoeffizienten der beiden Bedingungen „ME-logarithmisch -50 bis 100 Punkte“ und „ME-logarithmisch -100 bis 100 Punkte“ waren höher als der Koeffizient der Bedingung mit Multiple-Choice. Jedoch waren diese Differenzen nicht signifikant. Die Versuchsbedingung „ME-logarithmisch -300 bis 100 Punkte“ zeigte mit einem Z(r) von 0,48 (r=,47) eine signifikante Verringerung der Validität um 0,11 gegenüber der Bedingung mit Multiple-Choice.

Auch für schwierige Items nahm die Validität der Bedingungen mit Multipler Evaluation in Abhängigkeit von der Höhe der maximalen Strafzahlung in der Bedingung ab. Eine Ausnahme war die Bedingung „ME-logarithmisch -50 bis 100 Punkte“. Diese hatte mit einem Z(r) von 0,59 (r=,53) eine um 0,02 nicht signifikant höhere Validität als die Bedingung „ME-linear 0 bis 100 Punkte“ mit einem Z(r) von 0,57 (r=,52). Die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Fisher-Z-transformierten Korrelationskoeffizienten für schwierige Items können Tabelle 10-27 entnommen werden.

Tabelle 10-27

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen (nach Bortz, 2005) für die Daten der zwölf schwierigen Items (Stufe C).

Bedingung 1	Z(r)1	Bedingung 2	Z(r)2	**ΔZ(r)**	*z*	*df*	*p*
ME-linear 0 bis 100 Punkte	0,57	ME-logarithmisch -50 bis 100 Punkte	0,59	**0,02**	0,401	1	n.s.
		ME-logarithmisch -100 bis 100 Punkte	0,46	**-0,11***	-2,207	1	,05
		ME-logarithmisch -300 bis 100 Punkte	0,28	**-0,29****	-5,818	1	<,0005
ME-logarithmisch -50 bis 100 Punkte	0,59	ME-logarithmisch -100 bis 100 Punkte	0,46	**-0,13***	-2,608	1	,005
		ME-logarithmisch -300 bis 100 Punkte	0,28	**-0,31****	-6,219	1	<,0005
ME-logarithmisch -100 bis 100 Punkte	0,46	ME-logarithmisch -300 bis 100 Punkte	0,28	**-0,18****	-3,611	1	,0005
MC-0 oder 1 Punkt	0,47	ME-linear 0 bis 100 Punkte	0,57	**0,10***	2,006	1	,05
		ME-logarithmisch -50 bis 100 Punkte	0,59	**0,12****	2,407	1	,005
		ME-logarithmisch -100 bis 100 Punkte	0,46	**-0,01**	-0,201	1	n.s.
		ME-logarithmisch -300 bis 100 Punkte	0,28	**-0,19****	-3,812	1	,0005

Z(r) 1=Fisher-Z-transformierter Korrelationskoeffizient der ersten Bedingung des Einzelvergleichs, Z(r) 2=Fisher-Z-transformierter Korrelationskoeffizient der zweiten Bedingung, Δ Z(r)=Differenz zwischen Bedingung eins und zwei (Z(r)2-Z(r)1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Der Vergleich mit Multiple-Choice zeigte bei den schwierigen Items, dass die Validität der Bedingungen mit Multipler Evaluation mit der linearen Auswertung ohne Strafzahlung sowie der Auswertung mit einer maximalen Strafzahlung von -50 Punkten jeweils signifikant verbessert war. In der Bedingung „ME-logarithmisch -300 bis 100 Punkte“ verschlechterte sich hingegen der Korrelationskoeffizient signifikant um 0,19 im Vergleich zu der Bedingung mit Multiple-Choice.

Es kann zusammengefasst werden, dass die Bedingung mit Multipler Evaluation und einer linearen Auswertung die höchste Validität zeigte. Hingegen verschlechterte sich die Validität bei einer logarithmischen Auswertung umso mehr, je höher der Toleranzfaktor war, d.h. je höher die zu erwartenden Strafzahlungen waren. Beim Vergleich des Tests mit Multipler Evaluation mit dem Test mit Multiple-Choice fand sich eine verbesserte Validität nur mit der Auswertung durch die lineare Funktion. In den beiden Bedingungen mit einer maximalen Strafzahlung von -50 (T=0,5) bzw. -100 Punkten (T=1) war die Validität zwar verbessert, jedoch

nicht signifikant. Die Bedingung mit einem Auszahlungsbereich von -300 bis 100 Punkten (T=3) zeigte eine signifikante Verringerung der Validität.

10.3.6 Realismusindex

Auch im zweiten Experiment wurde ein individueller Realismusindex als Maß der Güte der Kalibrierung eines Testteilnehmers bestimmt. Der niedrigste Realismusindex betrug -2,78 und der höchste 1,59. Tabelle 10-28 zeigt die deskriptive Statistik des Realismusindex. Abbildung 10-5 zeigt den Realismusindex für die Versuchsbedingungen, unterteilt in zehn Perzentile. In einer einfaktoriellen Varianzanalyse über den Realismusindex mit dem Faktor „Auswertung" zeigten sich signifikante Unterschiede zwischen den Bedingungen ($F(3, 3228)=17{,}765$; $p<{,}001$; $\eta^2={,}016$). Die Prüfung auf Signifikanz der Differenzen zwischen den einzelnen Versuchsbedingungen erfolgte mithilfe von Scheffé-Tests. Je höher die Strafzahlungen in der Auswertung waren, desto unverfälschter berichteten die Teilnehmer ihre tatsächlichen subjektiven Antwortsicherheiten. Der höchste mittlere Realismusindex von 0,69 zeigte sich deshalb bei der Auswertung durch die logarithmische Funktion mit dem Auszahlungsbereich -300 bis 100 Punkte und der geringste mit 0,61 bei der Auswertung durch die lineare Funktion ohne Strafzahlung. Der mittlere

Tabelle 10-28

Deskriptive Statistik des Realismusindex.

Bedingung	Minimum	Maximum	**Mittel-wert**	Standard-fehler	Standard-abweichung	Median	N
ME-linear 0 bis 100 Punkte	-0,28	1,59	**0,61**	0,009	0,25	0,636	808
ME-logarithmisch -50 bis 100 Punkte	-0,13	1,12	**0,66**	0,008	0,21	0,686	808
ME- logarithmisch -100 bis 100 Punkte	-0,12	1,19	**0,67**	0,008	0,23	0,707	808
ME-logarithmisch -300 bis 100 Punkte	-2,78	1,27	**0,69**	0,010	0,27	0,747	808

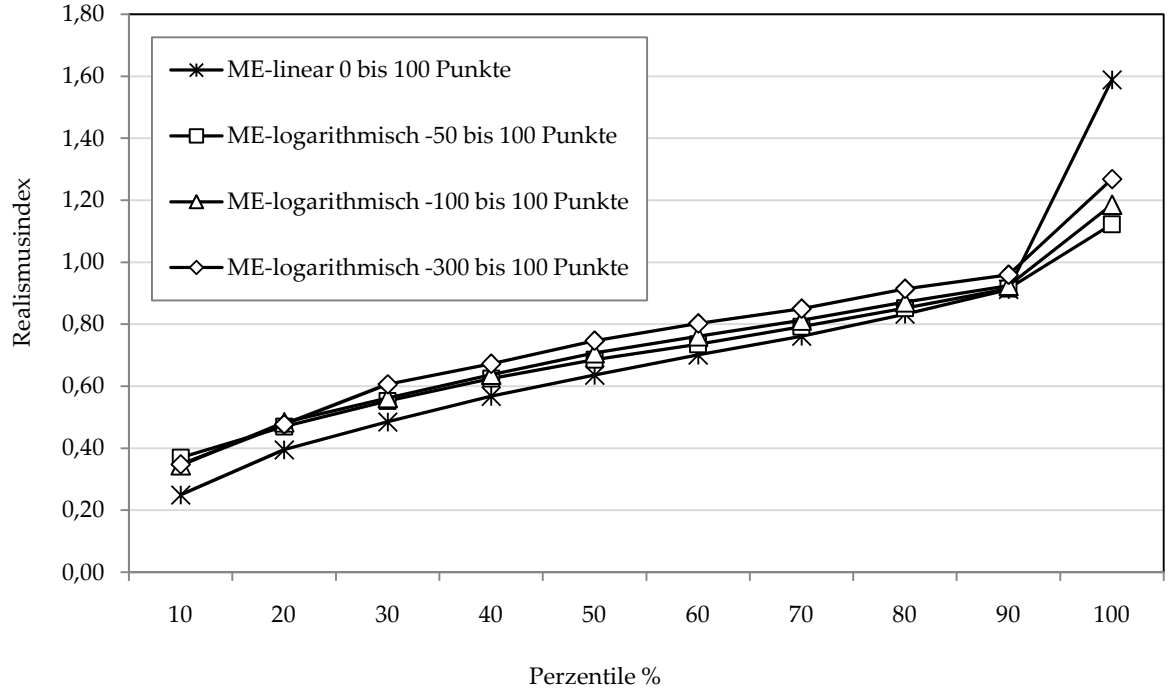

Abbildung 10-5: Realismusindex in den Versuchsbedingungen, dargestellt für zehn Perzentile.

Realismusindex war in allen Bedingungen, in denen die Punkte anhand der logarithmischen Funktion ausgezahlt wurden, signifikant höher als für die Bedingung, in der die Auswertung mithilfe der linearen Funktion erfolgte. Tabelle 10-29 zeigt die Differenzen zwischen den mittleren Realismusindices und die Ergebnisse der Prüfung auf Signifikanz mithilfe von Scheffé-Tests.

Tabelle 10-29

Realismusindexdifferenzen und Ergebnisse der Prüfung auf Signifikanz mithilfe von Scheffé-Tests.

	ME-logarithmisch -50 bis 100 Punkte		ME-logarithmisch -100 bis 100 Punkte		ME-logarithmisch -300 bis 100 Punkte	
	Differenz	*p*	Differenz	*p*	Differenz	*p*
ME-linear 0 bis 100 Punkte	0,05**	<,001	0,06**	<,001	0,08**	<,001
ME-logarithmisch -50 bis 100 Punkte			0,01	n.s.	0,03*	,044
ME-logarithmisch -100 bis 100 Punkte					0,02	n.s.

*=signifikant auf einem Niveau von 5%, **= signifikant auf einem Niveau von 1%.

10.3.7 Punktauszahlungen nach Realismuskorrektur

Bei der Realismuskorrektur wurden die Antwortsicherheiten in die richtige Antwort für jeden Testteilnehmer auf der Basis seines individuellen Realismusindex korrigiert (siehe Abschnitt 9.2.7). Diese Korrektur wurde beispielhaft anhand zweier Teilnehmer im Abschnitt 9.3.6 für die Auswertung durch die logarithmische Funktion nach Dirkzwager (2003) veranschaulicht. Da im zweiten Experiment auch die Auswertung durch eine lineare Funktion ohne Strafzahlungen untersucht wurde, wird nachfolgend auch für diese Auswertefunktion die Korrektur auf der Basis des Realismusindex für einen Teilnehmer, der *Overconfidence*, und für einen, der *Underconfidence* zeigte, verdeutlicht. Abbildung 10-6 zeigt die korrigierten und die nicht korrigierten Auszahlungen für alle Items eines Teilnehmers der mit einem Realismusindex von a=0,45 *Overconfidence* zeigt. Um die Korrektur grafisch zu veranschaulichen, wurden wiederum alle 36 Auszahlungen nach ihrer Höhe sortiert dargestellt. Bei der Korrektur wurden die Antwortsicherheiten in die richtige Antwort des Teilnehmers kleiner als 33% erhöht, beispielsweise bei 0% Antwortsicherheit auf 18%. Im Gegenzug wurden die Antwortsicherheiten

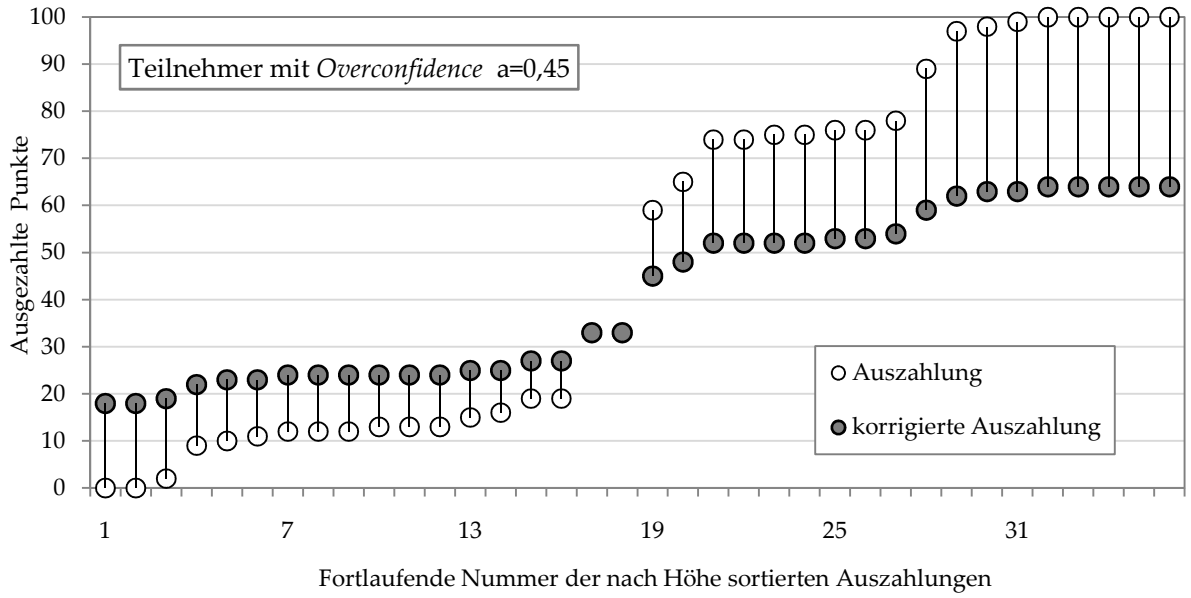

Abbildung 10-6: Korrigierte und nicht korrigierte Auszahlung eines Teilnehmers mit *Overconfidence* und einem Realismusindex von a=0,45. Die Auszahlungen werden in nach Höhe aufsteigend sortierter Reihenfolge dargestellt. Die Auswertung erfolgte mithilfe einer linearen Funktion mit einem Auszahlungsbereich von 0 bis 100 Punkten.

größer als 33% reduziert, beispielsweise von 100% auf 64%. Bei der Auswertung durch die lineare Funktion erhält der Teilnehmer seine prozentuale Antwortsicherheit in die richtige Antwort als Punkte ausgezahlt. Die korrigierten prozentualen Antwortsicherheiten wurden dem Teilnehmer deshalb als Punkte ausgezahlt. Gab der Teilnehmer mit 33% Antwortsicherheit völliges Nichtwissen wieder, so blieb diese Antwortsicherheit von 33% auch nach der Korrektur unverändert und der Teilnehmer erhielt erneut eine Auszahlung von 33 Punkten. Da es unter einer linearen Auswertung die beste Strategie für einen Teilnehmer ist, immer seine gesamte Antwortsicherheit in die plausibelste Antwortoption wiederzugeben, reduziert sich die Punktsumme des Teilnehmers durch die Korrektur. Der Teilnehmer, der mit einem Realismusindex von a=0,45 *Overconfidence* zeigte, erreichte vor der Korrektur eine Punktsumme von 1777 Punkten. Aufgrund der Korrektur erhält er 312 Punkte weniger und damit insgesamt nur noch 1465 Punkte ausgezahlt.

Um die Korrektur auf der Basis des Realismusindex bei der Auswertung durch die lineare Funktion für einen Teilnehmer, der *Underconfidence* zeigte, zu verdeutlichen, wird beispielhaft ein Teilnehmer mit einem Realismusindex von a=1,59 betrachtet. Abbildung 10-7 zeigt die korrigierten und die nicht korrigierten Auszahlungen des Teilnehmers für alle Items. Da der Teilnehmer *Underconfidence* zeigte, wurden seine Antwortsicherheiten in die richtige Antwort, die größer als 33% waren, durch die Korrektur erhöht, beispielsweise von 69% auf 90%. Die Antwortsicherheiten, die geringer als 33% waren, wurden dagegen reduziert, beispielsweise von 27% auf 23%. Da bei einer linearen Auswertung die berichteten Antwortsicherheiten in die richtige Antwort als Punkte ausgezahlt werden, erhält der Teilnehmer für Antwortsicherheiten kleiner als 33% eine geringere Punktauszahlung als vor der Korrektur. Da der Teilnehmer aber für seine Antwortsicherheiten in die richtige Antwort größer als 33% aufgrund der Korrektur deutlich mehr Punkte ausgezahlt bekam als ihm abgezogen wurden, betrug seine Punktsumme nach der Korrektur 1541 Punkte, und damit 128 Punkte mehr als vor der Korrektur, wo er 1431 Punkte erzielte.

Wurden alle Teilnehmer der Bedingungen mit der logarithmischen Auswertung betrachtet, so zeigte sich, dass sich die Punktsummen in den Bedingungen aufgrund der Korrektur einander angenähert hatten. Vor der Korrektur

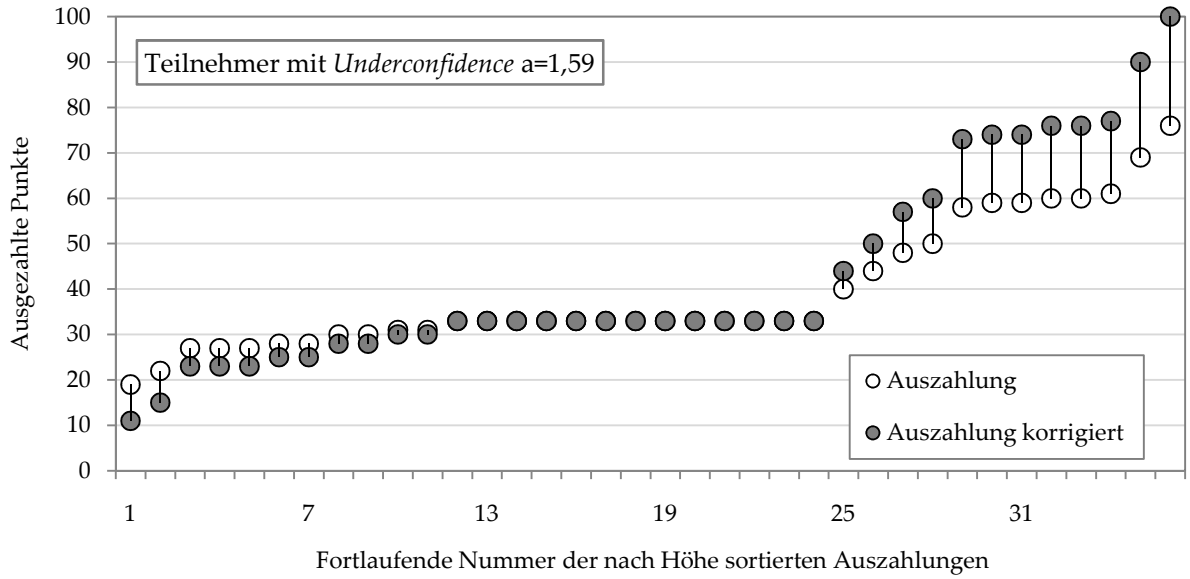

Abbildung 10-7: Korrigierte und nicht korrigierte Auszahlung eines Teilnehmers mit *Underconfidence* und einem Realismusindex von a=1,59. Die Auszahlungen werden in nach Höhe aufsteigend sortierter Reihenfolge dargestellt. Die Auswertung erfolgte mithilfe einer linearen Funktion mit einem Auszahlungs bereich von 0 bis 100 Punkten.

unterschieden sich die Punktsummen in den Bedingungen signifikant. Eine einfaktorielle Varianzanalyse für die drei Bedingungen, mit der Auswertung durch eine logarithmische Funktion, über die Punktsummen mit dem Faktor „Auswertung“ zeigte, dass sich die Punktsummen nach der Korrektur nicht mehr signifikant unterschieden ($F(2, 2421)=1{,}189$; $p<{,}305$; $\eta^2<{,}001$). Tabelle 10-30 zeigt die korrigierten und die nicht korrigierten Punktsummen des Englischtests. Eine Ausnahme stellte die lineare Funktion dar. Da die Funktion keine negativen Punkte auszahlte, wurde die Punktsumme insgesamt um 316,78 Punkte reduziert. Die korrigierte Punktsumme bei der linearen Auswertung war signifikant verschieden von den korrigierten Punktsummen in den Bedingungen mit Multipler Evaluation und der logarithmischen Auswertung. Dies zeigte eine einfaktorielle Varianzanalyse über die Punktsummen mit dem Faktor „Auswertung“ für die Stufen „lineare Auswertung“ und „logarithmische Auswertung“ ($F(1, 3230)=501{,}338$; $p<{,}001$; $\eta^2<{,}134$). Abbildung 10-8 zeigt die Mittelwerte der korrigierten und nicht korrigierten Punktauszahlung je Item für den gesamten Test und für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items. Die Tabellen der deskriptiven Statistik der

Punktsummen der realismuskorrigierten Auszahlungen aller 36 Items und der einfachen, mittelschwierigen und schwierigen Items für alle Versuchsbedingungen befinden sich im Anhang B.3.

Tabelle 10-30

Korrigierte und nicht korrigierte Punktsummen über alle 36 Items.

	korrigierte Punktsummen			nicht korrigierte Punktsummen		
Bedingungen	**Mittelwert**	Standard-abweichung	Standard-fehler	**Mittelwert**	Standard-abweichung	Standard-fehler
ME-linear 0 bis 100 Punkte	**2012,99**	607,84	21,38	**2329,77**	590,85	20,79
ME-logarithmisch -50 bis 100 Punkte	**1292,67**	852,84	30,00	**1754,62**	815,26	28,68
ME-logarithmisch -100 bis 100 Punkte	**1305,55**	870,36	30,62	**1463,61**	958,92	33,73
ME-logarithmisch -300 bis 100 Punkte	**1243,00**	860,18	30,26	**567,50**	1561,50	54,93

N=808 je Bedingung.

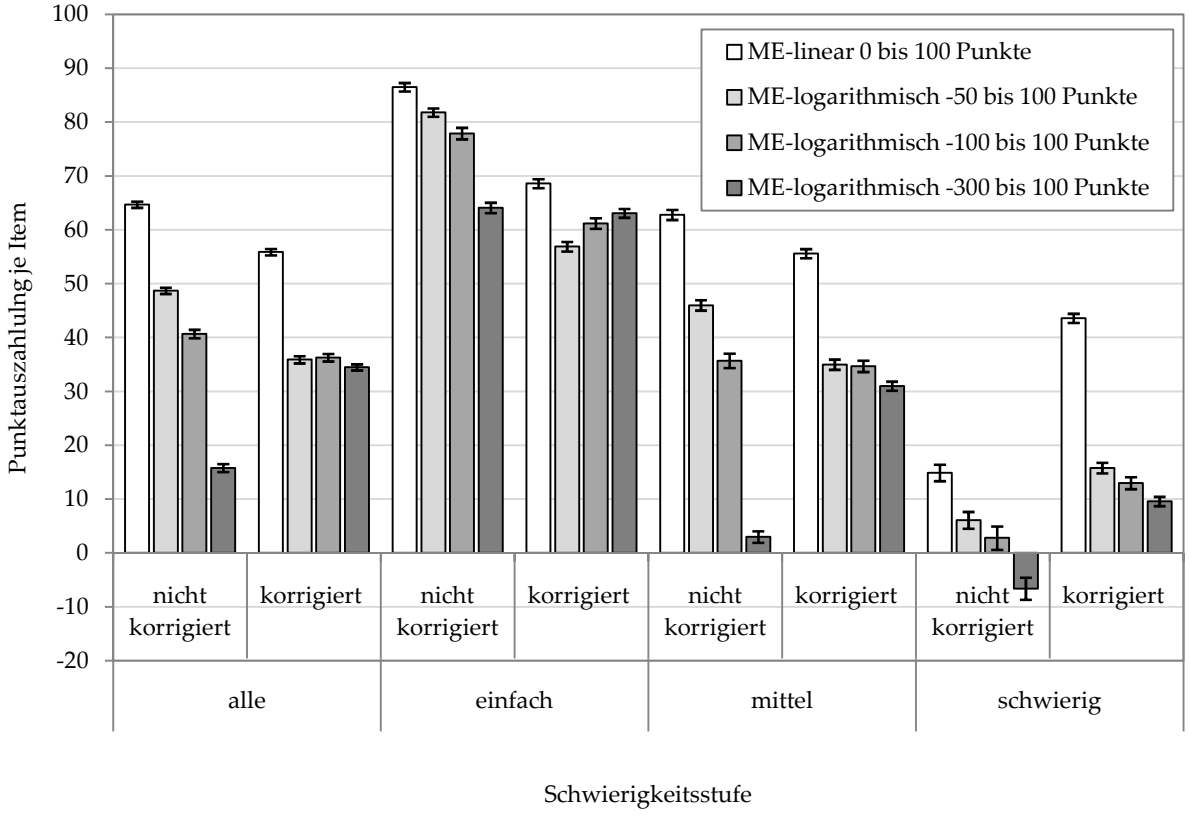

Abbildung 10-8: Mittelwert der Punktauszahlungen je Item für nicht korrigierte und korrigierte Punktauszahlungen. Die Fehlerbalken repräsentieren Standardfehler.

10.3.8 Reliabilität nach Realismuskorrektur

Für die anhand des Realismusindex der Teilnehmer korrigierten Punktauszahlungen wurden die Cronbach-α-Koeffizienten für alle 36 Items als Maß der Reliabilität berechnet. Tabelle 10-31 zeigt die α-Koeffizienten der korrigierten Auszahlungen. Die Cronbach-α-Koeffizienten der korrigierten Punktauszahlungen waren signifikant

Tabelle 10-31

Cronbach-α-Koeffizienten, berechnet für die auf der Basis des Realismusindex korrigierten Punktauszahlungen.

Schwierigkeitsstufe	Bedingung	α	α standardisierte Items	Anzahl Items	N
alle	ME-linear 0 bis 100 Punkte	0,96	0,96	36	808
	ME-logarithmisch -50 bis 100 Punkte	0,95	0,96	36	808
	ME-logarithmisch -100 bis 100 Punkte	0,93	0,93	36	808
	ME-logarithmisch -300 bis 100 Punkte	0,87	0,89	36	808
einfach (A)	ME-linear 0 bis 100 Punkte	0,96	0,97	12	808
	ME-logarithmisch -50 bis 100 Punkte	0,94	0,95	12	808
	ME-logarithmisch -100 bis 100 Punkte	0,92	0,93	12	808
	ME-logarithmisch -300 bis 100 Punkte	0,86	0,88	12	808
mittel (B)	ME-linear 0 bis 100 Punkte	0,90	0,90	12	808
	ME-logarithmisch -50 bis 100 Punkte	0,87	0,87	12	808
	ME-logarithmisch -100 bis 100 Punkte	0,82	0,83	12	808
	ME-logarithmisch -300 bis 100 Punkte	0,70	0,71	12	808
schwierig (C)	ME-linear 0 bis 100 Punkte	0,85	0,85	12	808
	ME-logarithmisch -50 bis 100 Punkte	0,84	0,84	12	808
	ME-logarithmisch -100 bis 100 Punkte	0,74	0,74	12	808
	ME-logarithmisch -300 bis 100 Punkte	0,59	0,60	12	808

höher als die α-Koeffizienten der nicht korrigierten. Die deutlichste Verbesserung um jeweils 0,08 zeigten die Bedingungen „ME-logarithmisch -50 bis 100 Punkte“ und „ME-logarithmisch -100 bis 100 Punkte“. Die Cronbach-α-Koeffizienten für nicht korrigierte und korrigierte Punkte und die Ergebnisse der Prüfungen auf Signifikanz der Differenzen zeigt Tabelle 10-32.

Für einfache Items waren die für die korrigierten Punkte errechneten α-Koeffizienten in allen Bedingungen signifikant verbessert. Die höchste Verbesserung um 0,22 zeigte sich bei der Auswertung durch die logarithmische

Tabelle 10-32

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten, ermittelt für korrigierte und nicht korrigierte Punktauszahlungen.

Schwierigkeits-stufe	Bedingung	α-nicht korri-giert	α-korri-giert	**$\Delta\ \alpha$**	X^2	*df*	*p*
alle	ME-linear 0 bis 100 Punkte	0,89	0,96	**0,07****	186,441	1	<,0001
	ME-logarithmisch -50 bis 100 Punkte	0,87	0,95	**0,08****	167,157	1	<,0001
	ME-logarithmisch -100 bis 100 Punkte	0,85	0,93	**0,08****	107,941	1	<,0001
	ME-logarithmisch -300 bis 100 Punkte	0,80	0,87	**0,07****	35,112	1	<,0001
einfach (A)	ME-linear 0 bis 100 Punkte	0,77	0,96	**0,19****	457,224	1	<,0001
	ME-logarithmisch -50 bis 100 Punkte	0,75	0,94	**0,19****	317,591	1	<,0001
	ME- logarithmisch -100 bis 100 Punkte	0,73	0,92	**0,19****	236,317	1	<,0001
	ME-logarithmisch -300 bis 100 Punkte	0,64	0,86	**0,22****	146,198	1	<,0001
mittel (B)	ME-linear 0 bis 100 Punkte	0,77	0,90	**0,13****	114,715	1	<,0001
	ME-logarithmisch -50 bis 100 Punkte	0,73	0,87	**0,14****	88,971	1	<,0001
	ME- logarithmisch -100 bis 100 Punkte	0,72	0,82	**0,10****	33,020	1	<,0001
	ME-logarithmisch -300 bis 100 Punkte	0,62	0,70	**0,08****	9,513	1	,0025
schwierig (C)	ME-linear 0 bis 100 Punkte	0,75	0,85	**0,10****	44,004	1	<,0001
	ME-logarithmisch -50 bis 100 Punkte	0,75	0,84	**0,09****	33,683	1	<,0001
	ME- logarithmisch -100 bis 100 Punkte	0,66	0,74	**0,08****	12,242	1	,0008
	ME-logarithmisch -300 bis 100 Punkte	0,67	0,59	**-0,08***	8,024	1	,0049

$\Delta\ \alpha$=Differenz zwischen α-korrigierte Punkte und α-nicht korrigierte Punkte (α korrigiert – α nicht korrigiert). Positive Differenzen bedeuten eine Verbesserung des korrigierten α, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%, N je Bedingung=808, Anzahl Items je Schwierigkeitsstufe=12, für alle Schwierigkeitsstufen 36.

Funktion mit dem Auszahlungsbereich von -300 bis 100 Punkten. Die drei weiteren Bedingungen mit Multipler Evaluation zeigten jeweils eine Verbesserung des Cronbach-α um 0,19.

Auch die Cronbach-α-Koeffizienten der mittelschwierigen Items erhöhten sich durch die Korrektur der Punktauszahlung auf der Basis des Realismusindex der Testteilnehmer signifikant. Die Verbesserungen der α-Koeffizienten waren jedoch etwas geringer als für einfache Items. Die höchste Verbesserung betrug 0,14 in der Bedingung „ME-logarithmisch -50 bis 100 Punkte“ und die geringste 0,08 in der Bedingung „ME-logarithmisch -300 bis 100 Punkte“.

Die Cronbach-α-Koeffizienten für schwierige Items waren ebenfalls signifikant verbessert, wenn die Punktauszahlungen korrigiert wurden. Eine Ausnahme war dabei die Bedingung mit der logarithmischen Auswertung mit dem Auszahlungsbereich von -300 bis 100 Punkten. In dieser Bedingung verschlechterte sich α signifikant um 0,08.

Auch im zweiten Experiment wurde für alle Versuchsbedingungen eine deutliche signifikante Verbesserung der Reliabilität aufgrund der Korrektur mithilfe des individuellen Realismusindex gezeigt.

10.3.9 Validität nach Realismuskorrektur

Um die Fragestellung zu untersuchen, ob die Korrektur der Antwortsicherheiten auf der Basis des Realismusindex zu einer Verbesserung der Validität führt, wurden die Korrelationen (nach Pearson) des z-Index der Selbsteinschätzung mit den Punktsummen der korrigierten Punkte ermittelt. Um eine Prüfung auf signifikante Unterschiede (Bortz, 2005) vornehmen zu können, wurden die Korrelationskoeffizienten Fisher-Z-transformiert. Tabelle 10-33 zeigt die Korrelationskoeffizienten und die Z-transformierten Koeffizienten, ermittelt für alle Items der Bedingungen mit Multipler Evaluation. Für den gesamten Test zeigte sich

Tabelle 10-33

Korrelationskoeffizienten und Fisher-Z-transformierte Koeffizienten des z-Index der Selbsteinschätzung mit den auf der Basis der realismuskorrigierten Antwortsicherheiten ausgezahlten Punktsummen aller 36 Items.

Bedingung	r	Z(r)	N
ME-linear 0 bis 100 Punkte	,64	0,75	808
ME-logarithmisch -50 bis 100 Punkte	,61	0,70	808
ME-logarithmisch -100 bis 100 Punkte	,59	0,67	808
ME-logarithmisch -300 bis 100 Punkte	,62	0,73	808

eine signifikante Erhöhung der Validität von Z(r)=0,54 (r=,50) um 0,19 auf Z(r)=0,73 (r=,62) nur für die Bedingung, in der die Punkte anhand der logarithmischen Funktion mit dem Auszahlungsbereich von -300 bis 100 Punkte ausgezahlt wurden. Die anderen Bedingungen zeigten dagegen geringe, nicht signifikante

Verringerungen der Validität. Tabelle 10-35 zeigt die Ergebnisse der Prüfungen auf signifikante Unterschiede zwischen den Korrelationskoeffizienten.

Wurde die Validität nur für die einfachen Items betrachtet, so zeigten alle Versuchsbedingungen eine geringe Verbesserung der Validität aufgrund der Korrektur. Jedoch war keine dieser Verbesserungen signifikant, mit Ausnahme der

Tabelle 10-34

Korrelationskoeffizienten und Fisher-Z-transformierte Koeffizienten des z-Index der Selbsteinschätzung mit den auf der Basis der realismuskorrigierten Antwortsicherheiten ausgezahlten Punktsummen der jeweils zwölf einfachen, mittelschwierigen und schwierigen Items.

Schwierigkeitsstufe	einfach (A)		mittel (B)		schwierig (C)	
Bedingung	r	Z(r)	r	Z(r)	r	Z(r)
ME-linear 0 bis 100 Punkte	,62	0,72	,63	0,73	,52	0,58
ME-logarithmisch -50 bis 100 Punkte	,57	0,65	,57	0,65	,53	0,59
ME-logarithmisch -100 bis 100 Punkte	,54	0,60	,56	0,64	,46	0,50
ME-logarithmisch -300 bis 100 Punkte	,55	0,62	,56	0,63	,47	0,51

Verbesserung in der Bedingung „ME-logarithmisch -300 bis 100 Punkte“. Tabelle 10-34 zeigt die Korrelationskoeffizienten und die Z-transformierten Koeffizienten, ermittelt für die jeweils zwölf Items der drei Schwierigkeitsstufen. Die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Bedingungen können für einfache, mittelschwierige und schwierige Items ebenfalls der Tabelle 10-35 entnommen werden.

Wurden ausschließlich die mittelschwierigen Items betrachtet, so wurde eine signifikante Verbesserung der Validität durch die Realismuskorrektur wiederum nur für die Bedingung „ME-logarithmisch -300 bis 100 Punkte“ ermittelt. Dieses Befundmuster zeigte sich auch für die schwierigen Items. Nur die Validität, ermittelt für die Bedingung mit einer maximalen Strafzahlung von -300 Punkten, zeigte eine signifikante Verbesserung aufgrund der Realismuskorrektur.

Zusammenfassend wurde eine signifikante Verbesserung der Validität aufgrund der Realismuskorrektur nur für die Bedingung mit der logarithmischen Auswertung mit dem Toleranzfaktor T=3, die also hohe Strafzahlungen beinhaltet, gezeigt.

Tabelle 10-35

Ergebnisse der Prüfung auf signifikante Unterschiede (nach Bortz, 2005) zwischen den Korrelationskoeffizienten der korrigierten und nicht korrigierten Fisher-Z-transformierten Korrelationskoeffizienten.

Schwierigkeits-stufe	Bedingung	Z(r) nicht korrigiert	Z(r) korrigiert	**Δ Z(r)**	z	*df*	*p*
alle	ME-linear 0 bis 100 Punkte	0,81	0,75	**-0,06**	-1,195	1	n.s.
	ME-logarithmisch -50 bis 100 Punkte	0,78	0,70	**-0,08**	-1,595	1	n.s.
	ME-logarithmisch -100 bis 100 Punkte	0,70	0,67	**-0,02**	-0,435	1	n.s.
	ME-logarithmisch -300 bis 100 Punkte	0,54	0,73	**0,19****	3,763	1	,0005
einfach (A)	ME-linear 0 bis 100 Punkte	0,66	0,72	**0,06**	1,187	1	n.s.
	ME-logarithmisch -50 bis 100 Punkte	0,64	0,65	**0,01**	0,237	1	n.s.
	ME-logarithmisch -100 bis 100 Punkte	0,56	0,60	**0,04**	0,884	1	n.s.
	ME-logarithmisch -300 bis 100 Punkte	0,53	0,62	**0,09***	1,815	1	,05
mittel (B)	ME-linear 0 bis 100 Punkte	0,73	0,73	**0**	-0,033	1	n.s.
	ME-logarithmisch -50 bis 100 Punkte	0,66	0,65	**-0,01**	-0,210	1	n.s.
	ME-logarithmisch -100 bis 100 Punkte	0,63	0,64	**0,01**	0,175	1	n.s.
	ME-logarithmisch -300 bis 100 Punkte	0,48	0,63	**0,15****	3,089	1	,005
schwierig (C)	ME-linear 0 bis 100 Punkte	0,57	0,58	**0,01**	0,220	1	n.s.
	ME-logarithmisch -50 bis 100 Punkte	0,59	0,59	**0**	0,028	1	n.s.
	ME-logarithmisch -100 bis 100 Punkte	0,46	0,50	**0,04**	0,877	1	n.s.
	ME-logarithmisch -300 bis 100 Punkte	0,28	0,51	**0,23****	4,602	1	,0005

Z(r)=Fisher-Z-transformierte Korrelationskoeffizienten,
Δ Z(r)=Differenz zwischen Z(r) korrigierte Punkte und Z(r) nicht korrigierte Punkte,
*=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

10.3.10 Abbruchquoten

Auch im zweiten Experiment beendeten manche Testteilnehmer ihre Teilnahme am Experiment vorzeitig. Die im Folgenden als Abbruchquoten bezeichneten Prozentwerte bedeuten, dass von 100% der Teilnehmer, die den jeweiligen Teil des Experiments aufgerufen hatten, der dargestellte Prozentsatz von Teilnehmern nicht zum nächsten Teil des Experiments weiterging, sondern das Experiment an dieser Stelle abbrach. Der erste Teil der Testapplikation, der Fragebogen zur Erhebung der demographischen Daten, war für alle fünf Versuchsbedingungen gleich gestaltet.

Während der Bearbeitung dieses Fragebogens beendeten durchschnittlich 11,48% (*SD*=0,31%) der Teilnehmer das Experiment. Während der Einführung in das Antwortverfahren brachen in den Bedingungen mit Multipler Evaluation im Durchschnitt über die vier Bedingungen 2,3% (*SD*=0,28%) der Teilnehmer das Experiment ab, in der Bedingung mit Multiple-Choice 0,58%. In den Bedingungen mit Multipler Evaluation führten die Teilnehmer fünf Übungsaufgaben durch. Während dieser Übungen beendeten im Mittel 14,36% (*SD*=1,20%) der Teilnehmer das Experiment. Die Teilnehmer der Bedingung mit Multiple-Choice bearbeiteten keine Übungen.

Deutliche Unterschiede zwischen den Abbruchquoten zeigten sich während der Bearbeitung des Englischtests. Abbildung 10-9 zeigt die Abbruchquoten in Prozent während der Bearbeitung der Items des Tests. Die geringste Quote zeigte mit 13,02%

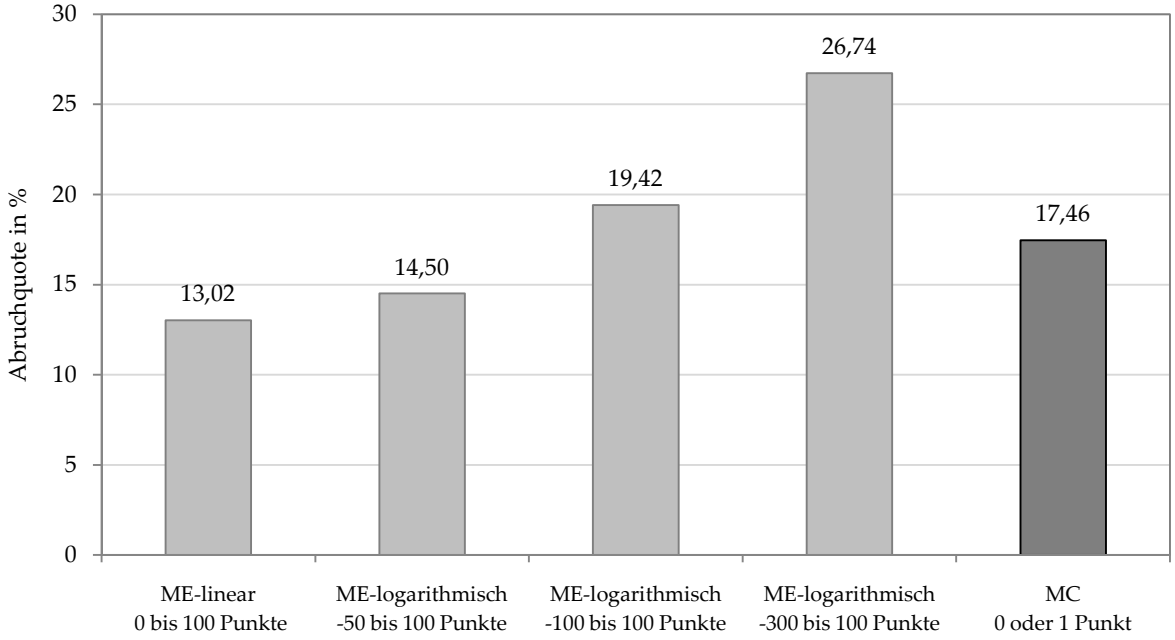

Abbildung 10-9: Abbruchquoten in Prozent während der Bearbeitung der Items des Englischtests.

die Bedingung mit Multipler Evaluation und der Auswertung durch die lineare Funktion. Je höher die Strafzahlungen in den Bedingungen mit einer logarithmischen Auswertung waren, desto mehr Testteilnehmer beendeten das Experiment während

der Bearbeitung der Items vorzeitig. So war die Abbruchrate um 11,39% höher, wenn die maximal mögliche Strafzahlung in der Bedingung -50 Punkte betrug (Δ=1,48%). Um 49,14% höher als mit der linearen Auswertung ohne Strafzahlung war die Abbruchquote, wenn die mögliche Strafzahlung in der Bedingung -100 Punkte betrug (Δ=6,40%), und bei einer maximalen möglichen Strafzahlung von -300 Punkten war die Abbruchquote um 105,34% höher als in der Bedingung mit der linearen Auswertung (Δ=13,72%). In der Bedingung mit Multiple-Choice beendeten 34,14% der Teilnehmer mehr die Bearbeitung der Items vorzeitig als bei der Auswertung mithilfe der linearen Funktion (Δ=4,44%).

Zusammenfassend zeigte sich während des Englischtests mit Multipler Evaluation eine Erhöhung der Abbruchquote in Abhängigkeit von der Höhe der Strafzahlungen in der Auswertung.

10.4 Diskussion

Im zweiten Experiment wurde der Einfluss der Höhe des Toleranzfaktors und damit der maximalen Strafzahlungen in der logarithmischen Auswertung (Dirkzwager, 2003) auf die Reliabilität und die Validität des Englischtests untersucht. Dazu erfolgte die Auszahlung der Punkte mit den Toleranzfaktoren T=0,5, T=1 bzw. T=3, d.h. einer maximalen Strafzahlung von -50, -100 bzw. -300 Punkten. Eine Bedingung ohne Strafzahlungen wurde durch eine lineare Funktion operationalisiert. Dabei wurde wiederum das Antwortverfahren Multiple Evaluation im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren betrachtet.

Die Hypothese, die Bedingungen mit Multipler Evaluation und einer logarithmischen Auswertung würden eine Verbesserung der Reliabilität in Form der internen Konsistenz in Abhängigkeit von der Höhe des Toleranzfaktors bzw. der Strafzahlungen zeigen, wurde nicht bestätigt. Entgegen der Erwartung verschlechterte sich stattdessen die Reliabilität des Englischtests umso mehr, je höher die Strafzahlungen waren. Der höchste Cronbach-α-Koeffizient wurde stattdessen beobachtet, wenn die Punkte durch die lineare Funktion, also ohne Strafzahlungen, ausgezahlt wurden und der niedrigste bei einer logarithmischen Auswertung mit dem Toleranzfaktor T=3, also einer Strafzahlung von maximal -300 Punkten. Dieser Effekt zeigte sich für den gesamten Englischtest und auch wenn jeweils die einfachen, mittelschwierigen oder schwierigen Items getrennt betrachtet wurden.

Die Reliabilität des Tests mit Multipler Evaluation war deshalb im Vergleich zum Test mit Multiple-Choice am deutlichsten verbessert, wenn die Punkte anhand der linearen Funktion ausgezahlt wurden. Um die interne Konsistenz dieser Bedingung mit Multipler Evaluation zu erreichen, müsste der Test mit Multiple-Choice um den Faktor 1,66 verlängert werden. Die Umrechnung in den Faktor Testverlängerung erfolgte anhand der Methode von Spearman-Brown (Bühner, 2006). Wenn die maximalen Strafzahlungen bei der Auswertung durch die logarithmische Funktion (Dirkzwager, 2003) -50 Punkte betrugen (T=0,5), war die Reliabilität des Tests mit Multipler Evaluation ebenfalls gegenüber der des Tests mit Multiple-Choice verbessert. Diese Verbesserung entsprach einer Verlängerung des Tests mit Multiple-Choice um den Faktor 1,37. Bei der Auswertung mit einer maximalen Strafzahlung von -100 Punkten (T=1) war die Reliabilität zwar auch noch höher als die der Bedingung mit Multiple-Choice, jedoch nicht signifikant. Hier

entsprach die Erhöhung des α-Koeffizienten einer Verlängerung des Tests mit Multiple-Choice um den Faktor 1,16. Die Reliabilität der Bedingung mit dem höchsten Toleranzfaktor (T=3), also einer hohen maximalen Strafzahlung von -300 Punkten, war hingegen signifikant gegenüber Multiple-Choice verschlechtert. Um die Reliabilität dieser Bedingung zu erreichen, würde ein um den Faktor 0,82 gekürzter Englischtest mit Multiple-Choice genügen.

Wurden jedoch ausschließlich die schwierigen Items betrachtet, so war die Reliabilität in allen Bedingungen mit Multipler Evaluation signifikant höher als in der Bedingung mit Multiple-Choice. Bei der Auswertung durch die lineare Funktion ohne Strafzahlungen sowie durch die logarithmische Funktion mit einer maximalen Strafzahlung von -50 Punkten entsprach diese Verbesserung einer Verlängerung des Multiple-Choice-Tests um den Faktor 2,00. Wurden die Punkte durch die logarithmischen Funktionen mit einer maximalen Strafzahlung von -100 bzw. -300 Punkten ausgezahlt, so entsprach die Verbesserung der Reliabilität einer Verlängerung des Multiple-Choice-Tests um den Faktor 1,29 bzw. 1,35.

Die Hypothese, dass sich die Validität des Englischtests verbessert, je höher die Strafzahlungen in der logarithmischen Auswertung sind, wurde anhand der Befunde ebenfalls nicht belegt. Stattdessen verringerte sich die Validität umso mehr, je höher die Strafzahlungen waren. Bei der Auswertung durch die lineare Funktion ohne Strafzahlungen zeigte sich die höchste Validität. Dieser Effekt zeigte sich für den gesamten Englischtest und auch wenn die einfachen, mittelschwierigen oder schwierigen Items getrennt betrachtet wurden.

Beim Vergleich der Multiplen Evaluation mit Multiple-Choice fand sich eine signifikant verbesserte Validität nur in den Bedingungen, in denen nur geringe Antwortsicherheiten in die richtige Antwortoption nicht durch hohen Punktabzug bestraft wurden. So zeigte sich mit der Auswertung durch die lineare Funktion und mit der logarithmischen Funktion mit einer maximalen Strafzahlung von -50 Punkten eine signifikant höhere Validität im Vergleich zu der Bedingung mit Multiple-Choice. Betrug die maximale Strafzahlung bei einer logarithmischen Auswertung -100 Punkte, dann war die Validität zwar im Vergleich zu der Bedingung mit Multiple-Choice verbessert, jedoch nicht signifikant. Die Bedingung mit einer logarithmischen Auswertung und einer Auszahlung von maximal

-300 Punkten zeigte eine signifikante Verringerung der Validität im Vergleich zu der Bedingung mit Multiple-Choice.

Bei schwierigen Items war die Validität in der Bedingung mit der linearen Auswertung sowie in der Bedingung mit der logarithmischen und einer maximalen Strafzahlung von -50 Punkten ebenfalls signifikant gegenüber der Bedingung mit Multiple-Choice verbessert. Bei der Auswertung durch die logarithmische Funktion mit hohen Strafzahlungen von bis zu -300 Punkten war bei schwierigen Items die Validität signifikant gegenüber Multiple-Choice verschlechtert.

Diese Ergebnisse stehen im Widerspruch zu den beispielsweise von Dirkzwager (2003), Shuford, Albert und Massengill (1966) und Shuford und Brown (1975) abgeleiteten Vorhersagen, dass die Auszahlung der Punkte anhand einer logarithmischen Funktion Teilnehmer inzentiviere, ihre tatsächlichen subjektiven Antwortsicherheiten unverfälscht wiederzugeben, und zwar umso mehr, je höher die maximalen Strafzahlungen sind. Als Folge daraus sollte die Varianz der beobachtbaren Testwerte reduziert und die Reliabilität und die Validität eines Tests verbessert werden können. Die Auswertung mithilfe einer linearen Funktion ohne Strafzahlung bewirkte stattdessen die höchste Reliabilität und Validität und damit auch die deutlichste Verbesserung der Güte des Tests im Vergleich zum herkömmlichen Multiple-Choice-Verfahren. Durch die Auszahlung der Punkte durch die logarithmische Funktion (Dirkzwager, 2003) verschlechterten sich die Reliabilität und die Validität im Vergleich dazu in Abhängigkeit von der Höhe der Strafzahlungen. Eine wesentliche Ursache dieser Befunde scheint in der nicht perfekten Kalibrierung vieler Teilnehmer zu liegen. Die Messung des Realismusindex als Maß der Güte der Kalibrierung zeigte, dass nur etwa ein Viertel der Teilnehmer eine annähernd perfekte Kalibrierung aufwiesen, während die übrigen Teilnehmer *Overconfidence* zeigten. Daher gaben diese Teilnehmer häufig hohe Antwortsicherheiten in die Distraktoren wieder und erhielten Strafpunkte. Die Höhe der Strafzahlungen zeigte deshalb auch einen signifikanten Einfluss auf die Punktsummen der Teilnehmer. Je höher die maximale Strafzahlung in der Bedingung war, desto niedriger waren die durchschnittlich erreichten Punktsummen. Daraus kann geschlossen werden, dass viele Teilnehmer ihr Antwortverhalten nicht aufgrund der hohen Strafzahlungen änderten. So konnte die Varianz der beobachtbaren Testwerte bei der Auswertung durch eine logarithmische

Funktion (Dirkzwager, 2003) nicht ausreichend reduziert werden, um die Reliabilität und die Validität deutlich zu verbessern. Stattdessen erhielten die Testteilnehmer teilweise hohe Strafzahlungen, die die psychometrische Qualität des Tests negativ beeinflussten. Die Abbruchquoten während des Englischtests lassen auch auf einen motivationalen Einfluss der Strafzahlungen auf die Teilnehmer schließen. In Abhängigkeit von der Höhe der möglichen Strafzahlungen erhöhte sich die Anzahl der Teilnehmer, die das Experiment abbrachen. Vermutlich frustrierten die hohen Strafzahlungen die Teilnehmer und sie beendeten deshalb ihre Teilnahme vorzeitig.

Die Ergebnisse des Experiments können dazu beitragen, die Inkonsistenz der bisherigen empirischen Befunde beispielsweise von Rippey (1968b) zu erklären. Rippey (1968b) fand sowohl Verbesserungen als auch Verschlechterungen der Reliabilität der Multiplen Evaluation im Vergleich zu Multiple-Choice. Die Reliabilität eines Tests mit Multipler Evaluation ist abhängig von der Art der verwendeten Auswertefunktion, aber auch von der Schwierigkeit des Testmaterials. So kann mit der Auswertung mit einem hohen Toleranzfaktor sowohl eine Verschlechterung der Reliabilität gegenüber Multiple-Choice gefunden werden, wenn die Items einfach oder mittelschwierig sind, aber ebenso eine Verbesserung, wenn die Items eine hohe Schwierigkeit aufweisen. Koele, De Boo und Verschure (1987) sowie Michael (1968) fanden eine Verbesserung der Reliabilität im Vergleich zu Multiple-Choice mit der Auswertung durch die lineare Funktion. Die Ergebnisse dieser Arbeit bestätigten diesen Befund. Auch Rippey (1970) fand die deutlichste Verbesserung der Reliabilität, wenn die Punkte durch eine lineare Funktion ausgezahlt wurden, während er mit der Auswertung durch eine logarithmische Funktion nach Shuford, Albert und Massengill (1966) nur eine geringfügige Verbesserung der Reliabilität gegenüber Multiple-Choice zeigen konnte. Die Annahme Rippeys (1970), dass diese Befunde in einer ungenügenden Information der Teilnehmer über die Punktauszahlung zu suchen seien, wurde durch die Ergebnisse dieser Arbeit nicht gestützt, denn es wurde im ersten Experiment gezeigt, dass eine Information über die Auszahlung keinen signifikanten Einfluss auf die Reliabilität ausübt. Die Ergebnisse des Experiments führen zu der Annahme, dass Hambleton, Roberts und Traub (1970), Koehler (1971) sowie Romberg, Shepler und Wilson (1970), die zur Auszahlung der Punkte eine logarithmische Funktion nach Shuford, Albert und Massengill (1966) verwendeten, keine Verbesserung der

Reliabilität gegenüber einem Multiple-Choice-Test zeigen konnten, weil ihr Testmaterial keinen genügend hohen Schwierigkeitsgrad aufwies.

Die Ergebnisse des zweiten Experiments zeigen aber auch, dass der Realismusindex der Teilnehmer durch die logarithmische Auswertung im Vergleich zur linearen zwar geringfügig, aber signifikant verbessert werden kann. Die Teilnehmer agierten bei der Wiedergabe ihrer Antwortsicherheiten vermutlich etwas vorsichtiger, um hohe Strafzahlungen zu vermeiden, und zwar umso mehr, je höher die mögliche zu erwartende Strafzahlung war. Deshalb kann gefolgert werden, dass Testteilnehmer durch hohe Strafzahlungen zumindest ansatzweise inzentiviert werden können, ihr Wissen unverfälscht zu reproduzieren. Das Ausmaß dieser Inzentivierung war aber zu gering, um die Reliabilität und die Validität des Tests zu verbessern.

Auch im zweiten Experiment wurde eine Korrektur der Antwortsicherheiten auf der Basis des individuellen Realismusindex der Teilnehmer vorgenommen. Dabei wurden die Antwortsicherheiten eines Teilnehmers nachträglich in die Antwortsicherheiten korrigiert, die er hätte wiedergeben müssen, wenn er perfekt kalibriert gewesen wäre. Die Hypothese, die Korrektur der Antwortsicherheiten mithilfe des individuellen Realismusindex würde zu einer Verbesserung der Reliabilität des Tests führen, konnte für alle Bedingungen mit Multipler Evaluation bestätigt werden. Nach dieser Korrektur waren alle Bedingungen mit Multipler Evaluation signifikant gegenüber der Bedingung mit Multiple-Choice verbessert. Die höchste Verbesserung der Reliabilität entsprach einer Verlängerung des Tests mit Multiple-Choice um den Faktor 4,92. Diese Verbesserung der Reliabilität wurde in der Bedingung mit der linearen Auswertefunktion beobachtet. Tabelle 10-36 zeigt

Tabelle 10-36

Faktoren, um die der Englischtest mit Multiple-Choice verlängert werden müsste, um die in der Bedingung mit Multipler Evaluation erzielte Reliabilität ebenfalls zu erreichen. Die Auszahlungen in den Bedingungen mit Multipler Evaluation wurden auf der Basis des Realismusindex korrigiert.

Bedingung	Faktor
ME-linear 0 bis 100 Punkte	4,92
ME-logarithmisch -50 bis 100 Punkte	3,89
ME-logarithmisch -100 bis 100 Punkte	2,72
ME-logarithmisch -300 bis 100 Punkte	1,36

für alle Bedingungen mit Multipler Evaluation die Faktoren, um die der Englischtest mit Multiple-Choice verlängert werden müsste, um die realismuskorrigierte Reliabilität zu erreichen.

Auch wenn die Cronbach-α-Koeffizienten der einfachen, mittelschwierigen und schwierigen Items getrennt betrachtet wurden, zeigte sich für alle Bedingungen eine signifikante Erhöhung von Cronbach-α. Die einzige Ausnahme stellte die Bedingung mit einer maximalen Strafzahlung von -300 Punkten bei schwierigen Items dar. Hier zeigte sich nach der Korrektur eine signifikante Verringerung von Cronbach-α. Zusammenfassend konnte damit für die interne Konsistenz gezeigt werden, dass diese Korrektur der Antwortsicherheiten mithilfe des individuellen Realismusindex eines Teilnehmers ein geeignetes Verfahren ist, um die Reliabilität eines Tests zu verbessern.

Die Validität konnte hingegen durch die Korrektur nicht verbessert werden. Eine Ausnahme war nur die Bedingung mit einer maximalen Strafzahlung von -300 Punkten. Bei dieser Auswertung wurde eine signifikante Verbesserung der Validität beobachtet. Die Hypothese, die auf der Basis der korrigierten Punktsummen ermittelte Validität sei höher als die der nicht korrigierten Punktsummen, konnte also nur für die Auswertung durch die logarithmische Funktion mit einem hohen Toleranzfaktor bestätigt werden. Eine mögliche Ursache dafür, dass ansonsten keine Verbesserung der Validität beobachtet wurde, könnte die Selbsteinschätzung als Maß der externen Validität sein. Möglicherweise ist diese selbsteingeschätzte Fähigkeit nicht sensitiv genug, um auch geringere Veränderungen der Varianz der beobachtbaren Testwerte eines Tests anzuzeigen.

11 Drittes Experiment

11.1 Fragestellungen und Hypothesen des dritten Experiments

Je geringer die Anzahl der Antwortoptionen eines Multiple-Choice-Tests ist, desto höher ist die Chance eines Teilnehmers, ein Item nur aufgrund von Raten richtig zu lösen. Daher sollte die Varianz der beobachtbaren Testwerte eines Tests, die durch das Raten der Teilnehmer entsteht, umso geringer sein, je größer die Anzahl der Antwortoptionen der Items gewählt wird. Die Anzahl der Antwortoptionen hat damit einen Einfluss auf die Güte eines Tests. Empirische Ergebnisse zeigten jedoch auch, dass es in der Regel nicht zu einer weiteren Verbesserung von Reliabilität und Validität führt, wenn die Anzahl der Antwortoptionen eines Items auf mehr als drei erhöht wird (Burton, 2001; Grier, 1975; Rodriguez, 2005). Einer der Gründe dafür ist beispielsweise, dass es häufig schwierig ist, genügend gleichwertige Distraktoren zu finden (Dressel & Schmid, 1953). Wird die Anzahl der Antwortoptionen eines Multiple-Choice-Tests aber von drei auf zwei reduziert, so ist, unter der Voraussetzung, dass beide Distraktoren gleichwertig sind, davon auszugehen, dass die Güte eines Tests dadurch verringert wird. Im dritten Experiment wurde diese Annahme geprüft. Es wurde beobachtet, ob die Reliabilität und die Validität eines Tests mit Multiple-Choice sich verschlechtern, wenn die Anzahl der Antwortoptionen von drei auf zwei verringert wird.

Bei einem Test mit Multipler Evaluation hat ein Teilnehmer hingegen die Möglichkeit, statt zu raten sein Nicht- und Teilwissen anzugeben. Daher sollte die Varianz in den beobachtbaren Testwerten und damit die Güte eines Tests nicht von der Anzahl der Antwortoptionen abhängig sein. Angesichts dessen wurde bei einem Test mit Multipler Evaluation aufgrund der Reduzierung der Anzahl der Antwortoptionen keine Verschlechterung der Reliabilität und der Validität erwartet. Diese Annahme wurde mithilfe des dritten Experiments empirisch überprüft.

Die Auswertung der Bedingungen mit Multipler Evaluation erfolgte mit einer logarithmischen Funktion (Dirkzwager, 2003) mit dem Toleranzfaktor T=3. Der Auszahlungsbereich reichte damit von -300 bis 100 Punkten. Bei dieser Auswertung erhielten die Teilnehmer also hohe Strafzahlungen, wenn sie nur eine geringe

Antwortsicherheit in die richtige Antwortoption angaben. Daher war es die einzig sinnvolle Strategie eines Teilnehmers, wenn er seine Gesamtpunktsumme maximieren wollte, seine tatsächlichen subjektiven Antwortsicherheiten vollkommen unverfälscht zu berichten (Shuford, Albert & Massengill, 1966).

Im zweiten Experiment zeigte der Test mit Multipler Evaluation jedoch mit der Auswertung durch eine lineare Funktion eine höhere Reliabilität und Validität als mit der Auswertung mit einer logarithmischen. Im dritten Experiment wurde deshalb auch diese nicht admissible Auswertung geprüft, bei der es für einen Teilnehmer die beste Strategie ist, immer der plausibelsten Antwortoption seine absolute Antwortsicherheit zuzuordnen. So wurde Multiple Evaluation mit einer reproduzierenden und mit einer nicht reproduzierenden Auswertebedingung im Vergleich zu Multiple-Choice untersucht.

Die Überprüfung der Fragestellungen erfolgte anhand der folgenden Hypothesen:

I. Hypothese:
Wenn die Anzahl der Antwortoptionen der Items von drei auf zwei reduziert wird, dann verringert sich die interne Konsistenz des Englischtests mit Multiple-Choice.

II. Hypothese:
Wenn die Anzahl der Antwortoptionen der Items von drei auf zwei reduziert wird, dann verringert sich die Validität des Englischtests mit Multiple-Choice.

III. Hypothese:
Wird die Anzahl der Antwortoptionen der Items von drei auf zwei reduziert, dann verringert sich die interne Konsistenz in den Bedingungen mit Multipler Evaluation nicht.

IV. Hypothese:
Wird die Anzahl der Antwortoptionen der Items von drei auf zwei reduziert, dann verringert sich die Validität in den Bedingungen mit Multipler Evaluation nicht.

Der Vergleich des Antwortverfahrens Multiple Evaluation mit dem Antwortverfahren Multiple-Choice erfolgte mithilfe der folgenden Hypothesen:

V. Hypothese:
Die Bedingungen mit Multipler Evaluation zeigen gegenüber den Bedingungen mit Multiple-Choice eine verbesserte interne Konsistenz.

VI. Hypothese:
Die Bedingungen mit Multipler Evaluation zeigen gegenüber den Bedingungen mit Multiple-Choice eine verbesserte Validität.

Bei einer logarithmischen Auswertung ist es die einzige Strategie für einen Teilnehmer, seine tatsächliche subjektive Antwortsicherheit völlig unverfälscht zu berichten, um sein Testergebnis zu maximieren. Bei einer linearen Auswertung ist es hingegen die beste Strategie eines Teilnehmers, immer der plausibelsten Antwortoption eine absolute Antwortsicherheit zuzuordnen. Daher war anzunehmen, dass die Varianz in den beobachtbaren Testwerten, die durch eine verfälschte Wiedergabe der tatsächlichen subjektiven Antwortsicherheiten entsteht, bei einer logarithmischen Auswertung im Vergleich zu einer linearen reduziert werden kann. Anhand der folgenden beiden Hypothesen wurde geprüft, ob deshalb die Reliabilität und die Validität eines Tests durch eine logarithmische Auswertung im Vergleich zu einer linearen verbessert werden können:

VII. Hypothese:
Der Englischtest mit Multipler Evaluation und einer logarithmischen Auswertung zeigt eine höhere Reliabilität als der Englischtest mit Multipler Evaluation mit einer linearen Auswertung.

VIII. Hypothese:
Der Englischtest mit Multipler Evaluation und einer logarithmischen Auswertung zeigt eine höhere Validität als der Englischtest mit Multipler Evaluation mit einer linearen Auswertung.

Auch im dritten Experiment wurde untersucht, ob eine Korrektur der Antwortsicherheiten mithilfe des individuellen Realismusindex (Holmes, 2002) ein geeignetes Verfahren ist, um die Reliabilität und die Validität eines Tests nachträglich signifikant zu verbessern. Dies wurde mithilfe der folgenden beiden Hypothesen geprüft:

IX. Hypothese:

Werden die Antwortsicherheiten der Testteilnehmer auf der Basis ihres individuellen Realismusindex korrigiert, so verbessert sich die interne Konsistenz, ermittelt auf der Basis der erneuten Punktauszahlung, im Vergleich zur internen Konsistenz, ermittelt für die Punkte, die mithilfe der nicht korrigierten Antwortsicherheiten ausgezahlt wurden.

X. Hypothese:

Werden die Antwortsicherheiten der Testteilnehmer auf der Basis ihres individuellen Realismusindex korrigiert, so verbessert sich die Validität, ermittelt auf der Basis der erneuten Punktauszahlung, im Vergleich zur Validität, ermittelt für die Punkte, die mithilfe der nicht korrigierten Antwortsicherheiten ausgezahlt wurden.

11.2 Methode

Im dritten Experiment wurden die Antwortsicherheiten wiederum durch Schieberegler erhoben, und die Teilnehmer aller Bedingungen erhielten eine Information über die Auszahlung. Die verwendete Testapplikation wurde auf der Basis der Applikationen des ersten und zweiten Experiments erstellt. Nachfolgend werden die Änderungen an der Applikation und in der Methode gegenüber dem ersten und dem zweiten Experiment beschrieben.

11.2.1 Design

Das Design des dritten Experiments war zweifaktoriell. Als erste unabhängige Variable wurde die Anzahl der Antwortoptionen in den Stufen „zwei Antwortoptionen“ und „drei Antwortoptionen“ untersucht. Für die Untersuchung der zweiten unabhängigen Variablen „Auswertefunktion“ erfolgte die Auszahlung der Punkte durch eine lineare Auswertefunktion mit einem Auszahlungsbereich von 0 bis 100 Punkten bzw. durch eine logarithmische nach Dirkzwager (2003) mit einem Auszahlungsbereich von -300 bis 100 Punkten (T=3). Die Kontrollbedingung war ein herkömmliches Multiple-Choice-Verfahren. Hierbei erhielten die Teilnehmer einen Punkt für eine richtige Antwort und null Punkte für eine falsche. Das Versuchsdesign ist in Tabelle 11-1 dargestellt.

Tabelle 11-1

2x3-faktorieller Versuchsplan.

Auswertefunktion	k=2	k=3
Multiple-Choice 0 oder 1 Punkt		
Multiple Evaluation linear 0 bis 100 Punkte		
Multiple Evaluation logarithmisch -300 bis 100 Punkte		

k=Anzahl der Antwortoptionen.

Um den Einfluss der Schwierigkeit der Items auf die Reliabilität und die Validität zu untersuchen, wurden die Ergebnisse für alle 36 Items des Tests sowie getrennt für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items, ermittelt.

11.2.2 Testmaterial

Als Testmaterial wurde derselbe Englischtest (siehe Abschnitt 8) verwendet wie im ersten und zweiten Experiment. Bei der Konstruktion der Items mit zwei Antwortoptionen sollten die Schwierigkeit und die Trennschärfe der Items so wenig wie möglich gegenüber denen der Items mit drei Antwortoptionen verändert werden. Deshalb wurde für die Items mit zwei Antwortoptionen derjenige der beiden Distraktoren beibehalten, den die Testteilnehmer in den Multiple-Choice-Bedingungen des ersten und zweiten Experiments im Durchschnitt am häufigsten gewählt hatten. Die Mittelwerte der richtigen Antworten in Prozent und die ausgewählten Distraktoren befinden sich im Anhang D.

11.2.3 Beschreibung der Stichprobe

Am dritten Experiment nahmen pro Versuchsbedingung 1100 Teilnehmer teil, insgesamt also 6600. Die Altersverteilung in den Bedingungen lag im Durchschnitt bei 29,1 Jahren (*SD*=11,4). Der jüngste Teilnehmer war 10 Jahre, da Teilnehmer unter 10 Jahren aus der Auswertung ausgeschlossen wurden, und der älteste 78 Jahre alt. Tabelle 11-2 zeigt die deskriptive Statistik der Altersverteilung in den Versuchsbedingungen.

Tabelle 11-2

Deskriptive Statistik der Altersverteilung in den Bedingungen.

Bedingung	Minimum	Maximum	Mittelwert	Standardfehler	Standardabweichung	Median	N
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	12	75	29,0	0,335	11,1	26,0	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	12	71	28,8	0,338	11,2	25,0	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	12	72	29,3	0,353	11,7	25,0	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	11	72	29,0	0,336	11,1	25,0	1100
MC-0 oder 1 Punkt 2 Antwortoptionen	10	75	29,6	0,351	11,7	26,0	1100
MC-0 oder 1 Punkt 3 Antwortoptionen	11	78	28,9	0,346	11,5	25,0	1100

Die Stichprobe bestand aus 62,2% weiblichen Teilnehmern (*SD*=0,8%) und 37,8% männlichen (*SD*=0,8%). Tabelle 11-3 zeigt die Anteile der Frauen und Männer in den Versuchsbedingungen in Prozent.

Tabelle 11-3

Anteile der Frauen und Männer in den Versuchsbedingungen in Prozent.

Bedingung	weiblich %	männlich %	N
ME-linear 0 bis 100 Punkte- 2 Antwortoptionen	60,8	39,2	1100
ME-linear 0 bis 100 Punkte- 3 Antwortoptionen	62,2	37,8	1100
ME-logarithmisch -300 bis 100 Punkte- 2 Antwortoptionen	62,1	37,9	1100
ME-logarithmisch -300 bis 100 Punkte- 3 Antwortoptionen	63,4	36,6	1100
MC-0 oder 1 Punkt- 2 Antwortoptionen	61,9	38,1	1100
MC-0 oder 1 Punkt- 3 Antwortoptionen	62,6	37,4	1100

11.2.4 Rekrutierung der Teilnehmer

Die Teilnehmer des dritten Experiments wurden durch die in Tabelle 11-4 aufgeführten Webseiten auf das Experiment hingewiesen. Die Tabelle zeigt nur die Webseiten, von denen aus mindestens 1% der Teilnehmer das Experiment aufriefen. Webseiten, die weniger Teilnehmer rekrutierten, wurden unter „Sonstige“ zusammengefasst.

Tabelle 11-4

Anteil vonTeilnehmern in Prozent, die von den Webseiten (Referrer) zum Englischtest geleitet wurden.

Internetadresse	Teilnehmer in Prozent
Sonstige	27,6%
http://www.uni-duesseldorf.de/	27,5%
http://www.google.de/ und google.com/	17,6%
http://www.englisch-lernen-im-internet.de/	8,8%
http://englisch-lernen-online.net/	5,1%
http://www.uni-wuerzburg.de/	3,2%
http://idw-online.de/	2,5%
http://www.testedich.de/	2,3%
http://www.lernen.sprachdirekt.de/	1,9%
http://www.zv.uni-wuerzburg.de/	1,3%
http://studien.allgemeine-psychologie.de/	1,2%
http://www.aerztezeitung.de/	1,0%

11.3 Ergebnisse

11.3.1 Schwierigkeitsstufen

Auch für das dritte Experiment wurde die Schwierigkeit der Items getrennt für die Bedingungen mit Multipler Evaluation und Multiple-Choice betrachtet. Denn in den Bedingungen mit Multipler Evaluation gaben die Testteilnehmer ihre prozentualen Antwortsicherheiten in alle Antwortoptionen wieder, während sie in den Bedingungen mit Multiple-Choice die am wahrscheinlichsten richtige Antwortoption auswählten. Je Schwierigkeitsstufe wurden dabei zwölf Items betrachtet. Bei den Items mit zwei Antwortoptionen ordneten die Testteilnehmer in den Bedingungen

Tabelle 11-5

Deskriptive Statistik der prozentualen Antwortsicherheit in die richtige Antwort für die Bedingungen mit Multipler Evaluation und Items mit drei und zwei Antwortoptionen.

		3 Antwortoptionen			2 Antwortoptionen				
GER-Stufe	Schwierigkeit	Mittelwert	Standardabweichung	**Mittelwert/ Item**	Mittelwert	Standardabweichung	**Mittelwert/ Item**	Anzahl Items	N
A	einfach	1071,72	165,32	**89,31**	1078,30	149,25	**89,86**	12	3232
B	mittel	794,53	251,04	**66,21**	880,02	193,72	**73,33**	12	3232
C	schwierig	565,96	218,58	**47,08**	721,11	197,55	**60,09**	12	3232
alle	alle	2431,2	541,87	**67,53**	2679,43	448,00	**74,43**	36	3232

In jeder Schwierigkeitsstufe konnten die Teilnehmer in der Summe maximal 1200% Antwortsicherheit erreichen, insgesamt maximal 3600%.

mit Multipler Evaluation der richtigen Antwortoption bei einfachen Items (GER-Stufe A) durchschnittlich je Item 89,86% Antwortsicherheit zu, bei mittelschwierigen (GER-Stufe B) 73,33% und bei schwierigen (GER-Stufe C) 60,09%. Tabelle 11-5 zeigt die deskriptive Statistik der prozentualen Antwortsicherheit in die richtige Antwort für die Bedingungen mit Multipler Evaluation und Items mit drei und zwei Antwortoptionen. In den Bedingungen mit Multipler Evaluation und Items mit drei Antwortoptionen ordneten die Testteilnehmer bei einfachen Items der richtigen Alternative durchschnittlich 89,31% Antwortsicherheit je Item zu, bei den mittelschwierigen Items 66,21% und bei den schwierigen 47,08%. Durch die Reduzierung der Anzahl der Antwortoptionen in den Bedingungen mit Multipler

Evaluation von drei auf zwei verringerte sich die Schwierigkeit der Items. Bei einfachen Items betrug diese Verringerung nur 0,61% (Δ=0,55%) Antwortsicherheit. Bei mittelschwierigen Items verringerte sich die Antwortsicherheit in die richtige Antwortoption um 9,71% (Δ=7,12%), bei schwierigen Items um 21,65% (Δ=13,01%). Für den gesamten Test betrachtet betrug die Verringerung der Antwortsicherheit in die richtige Antwort bei drei im Vergleich zu zwei Antwortoptionen für die Bedingungen mit Multipler Evaluation 9,27% (Δ=6,90%).

In der Bedingung mit Multiple-Choice und Items mit drei Antwortoptionen betrug der mittlere Schwierigkeitsindex (Bühner, 2006, S. 83) für einfache Items 90,40%, für mittelschwierige Items 74,21% und für schwierige Items 59,17%. Tabelle 11-6 zeigt die deskriptive Statistik der richtigen Antworten in Prozent für die Bedingungen mit Multiple-Choice und Items mit zwei und drei Antwortoptionen.

Tabelle 11-6

Prozent richtige Antworten für die Bedingungen mit Multiple-Choice und Items mit drei und zwei Antwortoptionen.

		3 Antwortoptionen			2 Antwortoptionen				
GER-Stufe	Schwierigkeit	Mittelwert	Standardabweichung	**Mittelwert/ Item**	Mittelwert	Standardabweichung	**Mittelwert/ Item**	Anzahl Items	N
A	einfach	1084,82	149,11	**90,40**	1074,91	167,32	**89,58**	12	1100
B	mittel	890,55	206,74	**74,21**	804,00	263,98	**67,00**	12	1100
C	schwierig	710,09	244,05	**59,17**	576,73	255,07	**47,98**	12	1100
alle	alle	2685,45	488,30	**75,60**	2454,64	568,42	**68,18**	36	1100

In jeder Schwierigkeitsstufe konnten die Testteilnehmer in der Summe maximal 1200% richtige Antworten erreichen, insgesamt maximal 3600%.

Der Schwierigkeitsindex von 89,58% verringerte sich durch die Reduzierung der Anzahl der Antwortoptionen von drei auf zwei für einfache Items nur um 0,91% (Δ=0,82%). Bei mittelschwierigen Items verringerte er sich um 9,72% (Δ=7,21%), bei schwierigen um 18,91% (Δ=11,19%). Wurden alle Items des Tests betrachtet, so verringerte sich die Schwierigkeit durch die Reduzierung der Anzahl der Antwortoptionen um 9,81% (Δ=7,42%).

11.3.2 Punktauszahlungen

Für die Bedingungen mit Multipler Evaluation betrug der Auszahlungsbereich 0 bis 100 Punkte bei der Auswertung durch die lineare Funktion und -300 bis 100 Punkte bei der Auswertung durch die logarithmische. Für die Betrachtung der unabhängigen Variablen „Anzahl der Antwortoptionen" wurden beide Auswertebedingungen zusammengefasst. In den Bedingungen mit Items mit drei Antwortoptionen erreichten die Testteilnehmer durchschnittlich 1626,66 Punkte. Bei den Items mit zwei Antwortoptionen waren es im Durchschnitt 1892,97 Punkte, also 266,31 Punkte mehr. Dieser Punktunterschied aufgrund der Reduzierung der Anzahl der Antwortoptionen von drei auf zwei war in einer zweifaktoriellen Varianzanalyse über die Punktsummen mit den Faktoren „Auswertefunktion" und „Anzahl der Antwortoptionen" signifikant ($F(1, 4399)=67{,}946$; $p<{,}001$; $\eta^2={,}015$). Die deskriptive Statistik der Punktsummen für alle Versuchsbedingungen zeigt Tabelle 11-7.

Tabelle 11-7

Deskriptive Statistik der Punktsummen über alle Items des Tests für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	**Mittelwert**	Standardfehler	Standardabweichung	Median	Mittelwert/ Item	Anzahl Items	N
ME-2 Antwortoptionen	-4015	3600	**1892,97**	26,58	1246,83	2235,00	52,58	36	2200
ME-3 Antwortoptionen	-6295	3600	**1626,66**	30,15	1414,35	1982,00	45,18	36	2200
ME-linear 0 bis 100 Punkte	1040	3600	**2552,75**	10,88	510,33	2553,50	70,91	36	2200
ME-logarithmisch -300 bis 100 Punkte	-6295	3600	**966,87**	30,69	1439,45	1075,50	26,86	36	2200
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	1300	3600	**2656,99**	13,77	456,57	2670,00	73,81	36	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	1040	3600	**2448,51**	16,26	539,36	2460,00	68,01	36	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-4015	3600	**1128,94**	39,70	1316,73	1143,50	31,36	36	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-6295	3600	**804,80**	46,31	1536,09	1005,50	22,36	36	1100

Die zweite unabhängige Variable war die verwendete Auswertefunktion. Für die Betrachtung dieses Faktors wurden jeweils die Punktauszahlungen der Items mit

zwei und drei Antwortoptionen zusammengefasst. Wurden die Punkte durch die lineare Funktion ausgezahlt, so betrug der Mittelwert der Punktsummen 2552,75 Punkte. Bei der Auswertung durch die logarithmische Funktion betrug die durchschnittliche Punktsumme der Teilnehmer 966,87 Punkte. Dieser Unterschied in Höhe von 1585,88 Punkten war deutlich signifikant, wie eine zweifaktorielle Varianzanalyse über die Punktsummen mit den Faktoren „Auswertefunktion" und „Anzahl der Antwortoptionen" für den Faktor „Auswertefunktion" zeigte (F(1, 4399)=2409,510; p<,001; η^2=,354). Eine signifikante Interaktion zwischen den Faktoren „Anzahl der Antwortoptionen" und „Auswertefunktion" zeigte sich dabei nicht (F(1, 4399)=3,204; p=,074; η^2=,010). Der signifikante Unterschied zwischen den Punktsummen aufgrund der Verwendung der linearen und der logarithmischen Auswertefunktion war zu erwarten, da die Auswertung der Punkte anhand zweier verschiedener Skalen erfolgte. Ein Teilnehmer erhielt beispielsweise in der Bedingung mit Items mit drei Antwortoptionen für 33% Antwortsicherheit in die richtige Antwort mit der linearen Auswertung 33 Punkte ausgezahlt. Bei der logarithmischen Auswertung erhielt er dafür 0 Punkte.

Die Schwierigkeit der Items hatte ebenfalls einen deutlichen signifikanten Einfluss auf die Punktsummen. Dies wurde in einer dreifaktoriellen Varianzanalyse über die Punktsummen mit den Faktoren „Auswertefunktion" und „Anzahl der Antwortoptionen" mit Messwiederholung auf dem dritten Faktor „Schwierigkeit" nach einer Greenhouse-Geisser-Korrektur gezeigt (F(1.994, 8766.043)=5022,644; p<,001; η^2=,533). Zudem bestand, ebenfalls nach einer Greenhouse-Geisser-Korrektur, eine signifikante Interaktion der Schwierigkeit der Items mit der Anzahl der Antwortoptionen (F(1.994, 8766.043)=66,817; p<,001; η^2=,015) sowie eine Interaktion der Schwierigkeit mit der verwendeten Auswertefunktion (F(1.994, 8766.043)=797,885; p<,001; η^2=,154). Auch zeigte sich eine geringe, aber signifikante Interaktion zwischen der Schwierigkeit der Items, der verwendeten Auswerte-funktion und der Anzahl der Antwortoptionen (F(1.994, 8766.043)=4,113; p=,016; η^2=,001). Abbildung 11-1 zeigt die Mittelwerte der Punktsummen für die Bedingungen mit zwei und drei Antwortoptionen. Abbildung 11-2 zeigt die Mittelwerte der Bedingungen für die unabhängige Variable „Auswertefunktion" mit den Stufen lineare und logarithmische Auswertung für einfache, mittelschwierige und schwierige Items.

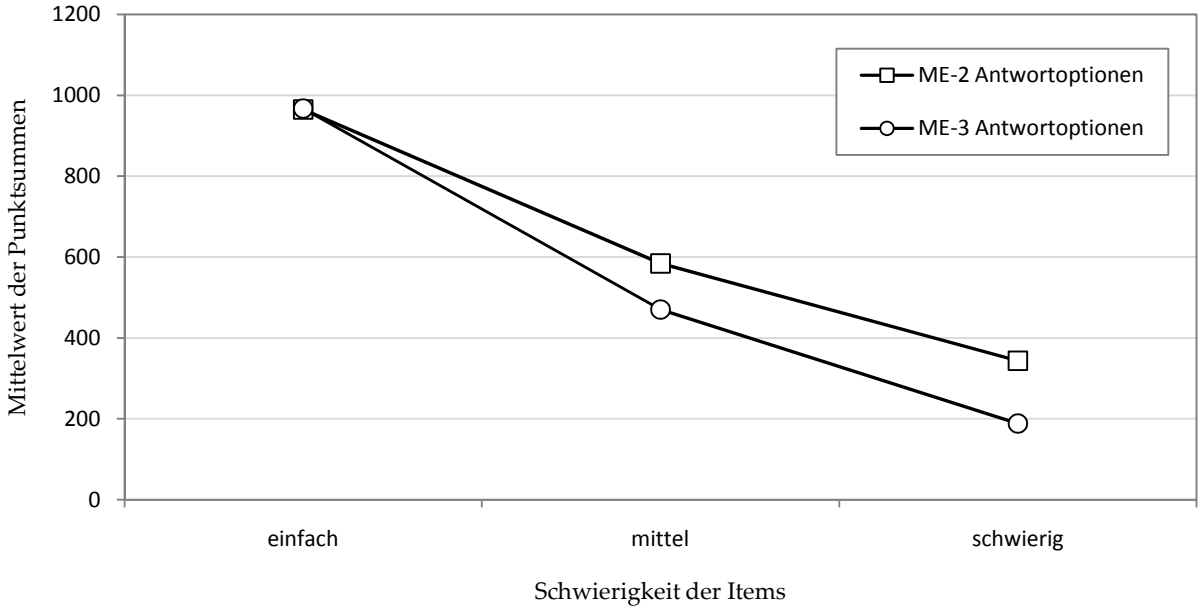

Abbildung 11-1: Mittelwerte der Punktsummen der Bedingungen mit zwei und drei Antwortoptionen. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

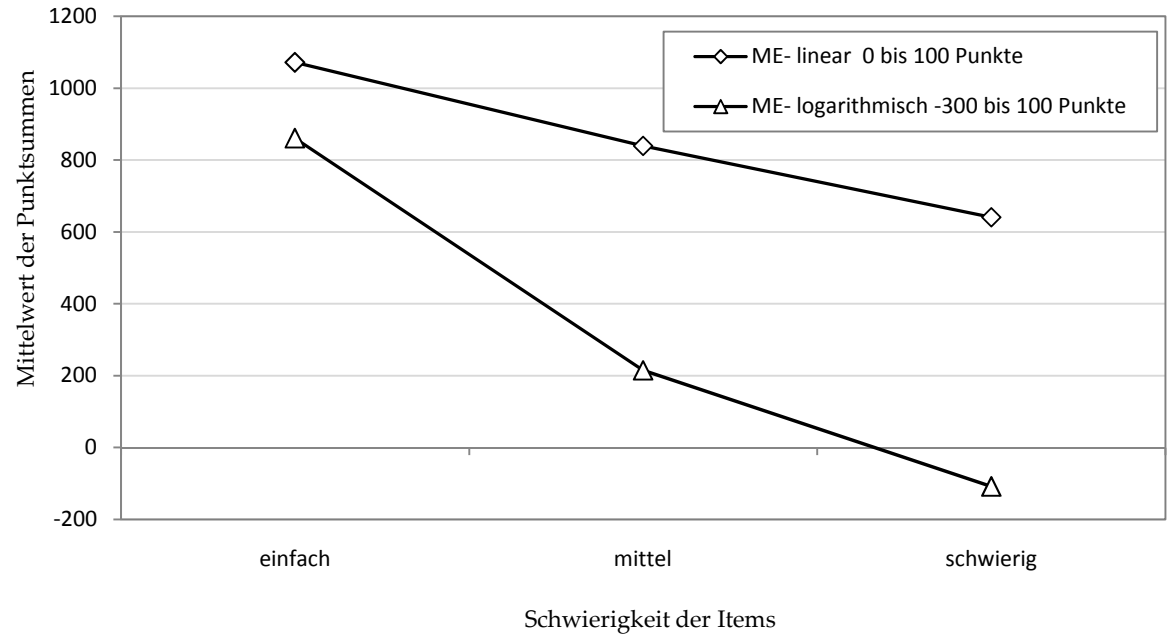

Abbildung 11-2: Mittelwerte der Punktsummen der Bedingungen mit der Auswertung durch die lineare und durch die logarithmische Funktion. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

In den Bedingungen mit Multiple-Choice erzielten die Testteilnehmer im Durchschnitt 26,85 Punkte für alle Items des Tests mit zwei Antwortoptionen. Für die Items mit drei Antwortoptionen betrug die mittlere Punktsumme 24,55 Punkte, und damit 2,4 Punkte weniger. Dieser Punktunterschied war signifikant, wie eine einfaktorielle Varianzanalyse über die Punktsummen mit dem Faktor „Anzahl der Antwortoptionen" zeigte (F(1, 2198)=104,364; p<,001; η^2<,045). Die deskriptive Statistik der Punktsummen für die Bedingungen mit Multiple-Choice für alle 36 Items des Tests zeigt Tabelle 11-8.

Tabelle 11-8

Deskriptive Statistik der Punktsummen für die Bedingungen mit Multiple-Choice für alle 36 Items des Tests.

Bedingung	Schwierig-keitsstufe	Mini-mum	Maxi-mum	**Mittel-wert**	Standard-abweich-ung	Standard-fehler	Median	Mittel-Wert/ Item
MC-2 Antwortoptionen	alle	12	36	**26,85**	4,88	0,147	27,00	0,746
MC-3 Antwortoptionen	alle	9	36	**24,55**	5,68	0,171	25,00	0,682

N je Bedingung=1100, Anzahl der Items=36.

Eine zweifaktorielle Varianzanalyse über die Punktsummen mit den Faktoren „Anzahl der Antwortoptionen" und einer Messwiederholung auf dem zweiten Faktor „Schwierigkeit" konnte nach einer Greenhouse-Geisser-Korrektur auch für die Bedingungen mit Multiple-Choice einen signifikanten Einfluss der Schwierigkeit der Items auf die Punktsummen belegen (F(1.917, 4215.521)=4241,924; p<,001; η^2=,659). Die Interaktion zwischen der Anzahl der Antwortoptionen und der Schwierigkeit der Items war ebenfalls, nach einer Greenhouse-Geisser-Korrektur, signifikant (F(1.917, 4215.521)=87,446; p<,001; η^2=,038). Abbildung 11-3 zeigt die Mittelwerte der Punktsummen der einfachen, mittelschwierigen und schwierigen Items für die Bedingungen mit Multiple Choice. Die deskriptive Statistik der Punktsummen für einfache, mittelschwierige und schwierige Items und die mittleren Punktauszahlungen je Item befinden sich in Anhang C.1.

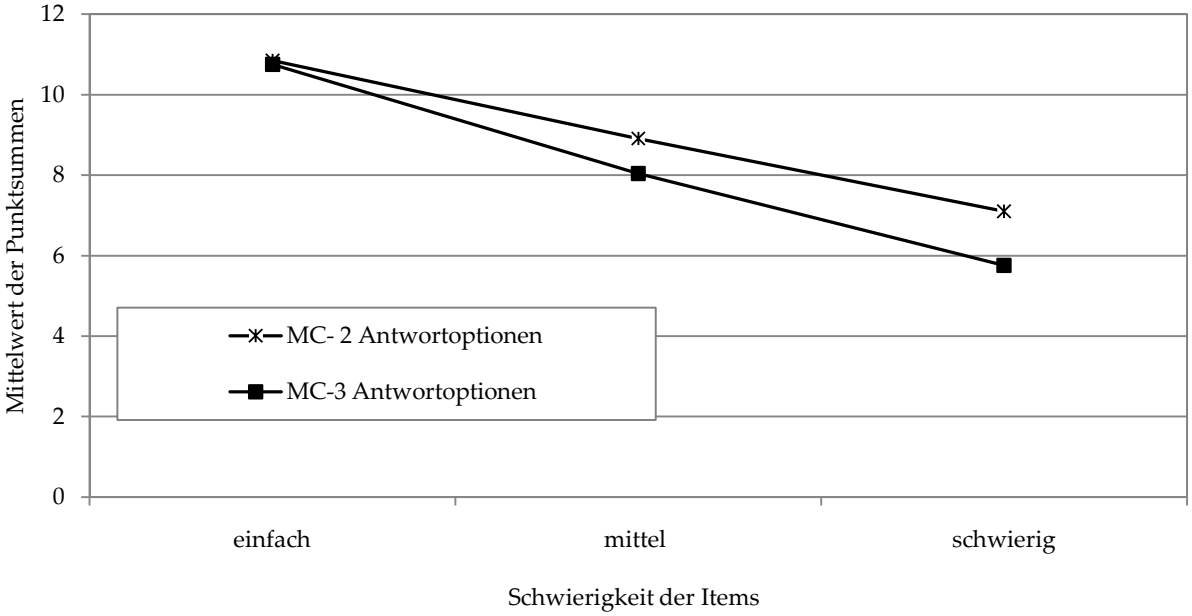

Abbildung 11-3: Mittelwerte der Punktsummen der Bedingungen mit Multiple-Choice. Die Standardfehler sind aufgrund der großen Stichprobe zu klein, um in der Abbildung sichtbar zu sein.

Wie dargelegt, unterschieden sich die Punktsummen in den Bedingungen mit Multipler Evaluation also signifikant, wenn die Auszahlung der Punkte anhand der linearen Funktion erfolgte, im Vergleich zu einer logarithmischen Auswertung. Auch die Punktsummen der Bedingungen mit Items mit zwei Antwortoptionen unterschieden sich signifikant von denen mit drei Antwortoptionen. Deshalb wurden die Bedingungen in den folgenden Auswertungen nicht für die unabhängigen Variablen „Auswertefunktion“ und „Anzahl der Antwortoptionen“ zusammengefasst betrachtet.

11.3.3 Trennschärfen

Für die ausgezahlten Punkte wurden die *part-whole-korrigierten* Trennschärfen ermittelt. Die niedrigste Trennschärfe von 0,09 hatte ein schwieriges Item mit zwei Antwortoptionen in der Bedingung mit Multiple-Choice. Die höchste Trennschärfe von 0,57 zeigte sich für ein schwieriges Item mit drei Antwortoptionen in der Bedingung mit Multipler Evaluation und der Auswertung durch die lineare Funktion. Die Trennschärfen aller Items je Bedingung befinden sich im Anhang C.2.

Um den Mittelwert der Trennschärfen bestimmen zu können, wurden alle Trennschärfen zunächst Fisher-Z-transformiert. Die Mittelwerte wurden einer inversen Fisher-Z-Transformation unterzogen. Tabelle 11-9 zeigt die Mittelwerte und Standardabweichungen der *part-whole-korrigierten* Trennschärfen, berechnet über alle 36 Items des Tests. Die höchste mittlere Trennschärfe von 0,37 zeigte sich in der Bedingung mit Multipler Evaluation, Items mit drei Antwortoptionen und der Auszahlung der Punkte durch die lineare Funktion. Bei der Auswertung durch die logarithmische Funktion und Items mit drei Antwortoptionen war die durchschnittliche Trennschärfe mit 0,30 um 0,07 geringer. Für die Items mit zwei Antwortoptionen war die mittlere Trennschärfe um 0,05 höher, wenn die Punkte durch die lineare Funktion ausgezahlt wurden im Vergleich zur Auswertung anhand

Tabelle 11-9

Mittelwerte der *part-whole-korrigierten* Trennschärfen.

Bedingung	Mittel-wert	Standard-abweichung	Anzahl Items	N
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	0,31	0,12	36	1100
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	0,37	0,10	36	1100
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	0,26	0,09	36	1100
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	0,30	0,07	36	1100
MC-2 Antwortoptionen	0,26	0,10	36	1100
MC-3 Antwortoptionen	0,31	0,09	36	1100

der logarithmischen. Für die Items mit zwei Antwortoptionen zeigte die Bedingung mit Multipler Evaluation und der linearen Auswertung mit einem Wert von 0,31 eine um 0,06 höhere Trennschärfe als die Bedingung mit Multiple-Choice mit einem Wert von 0,26. Die Bedingung mit Multipler Evaluation, Items mit zwei Antwortoptionen und der logarithmischen Auswertung zeigte eine gleich hohe mittlere Trennschärfe wie die Bedingung mit Multiple-Choice und Items mit zwei Antwortoptionen. Für die Items mit drei Antwortoptionen fand sich eine Verbesserung der durchschnittlichen Trennschärfe um 0,06 in der Bedingung mit Multipler Evaluation mit der Auswertung durch die lineare Funktion im Vergleich zu der Bedingung mit Multiple-Choice und Items mit drei Antwortoptionen. Die durchschnittliche Trennschärfe der Bedingung mit Multipler Evaluation, Items mit drei Antwort-

optionen und der Auswertung durch die logarithmische Funktion war dagegen um 0,01 niedriger als die in der Bedingung mit Multiple-Choice und Items mit drei Antwortoptionen.

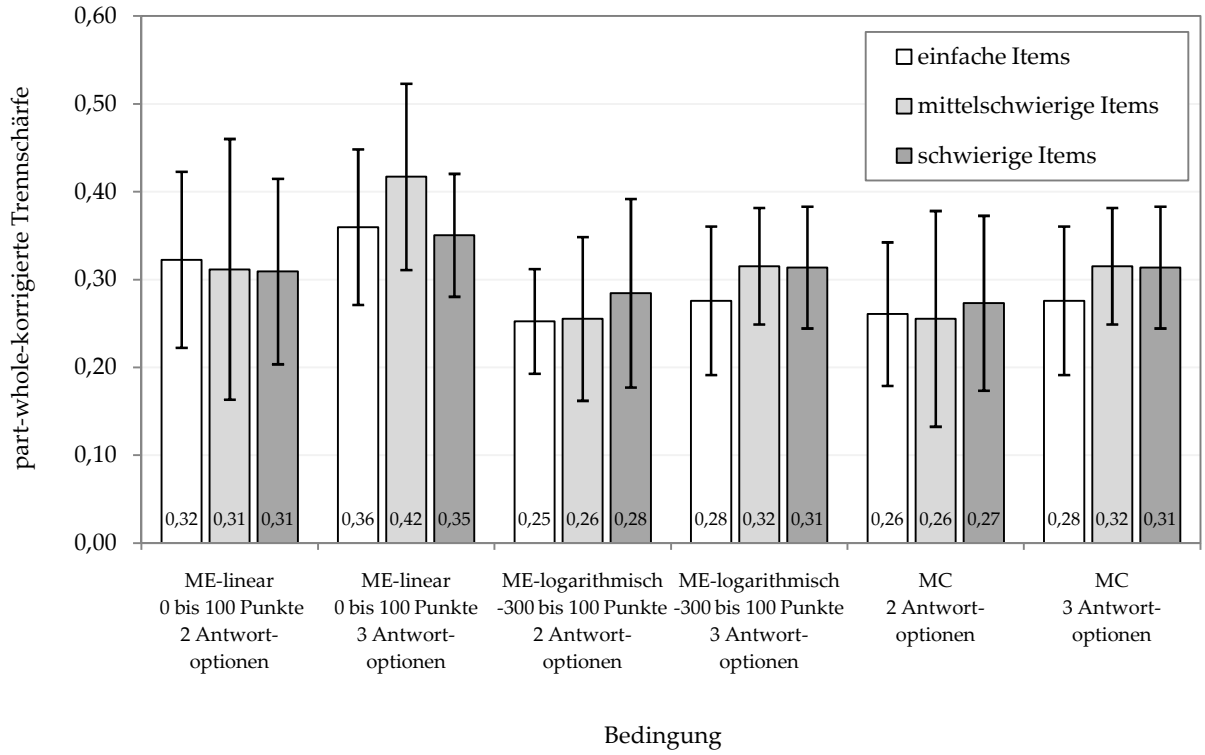

Abbildung 11-4: Mittelwerte der *part-whole-korrigierten* Trennschärfen für einfache, mittelschwierige und schwierige Items. Die Fehlerbalken repräsentieren Standardabweichungen.

Abbildung 11-4 zeigt die mittleren Trennschärfen für einfache, mittelschwierige und schwierige Items. Die deutlichsten Verbesserungen der Trennschärfen, im Vergleich der Items mit drei und zwei Antwortoptionen, zeigten sich sowohl für einfache und mittelschwierige, als auch für schwierige Items, mit der Auswertung durch die lineare Funktion und Items mit drei Antwortoptionen. Bei der Auswertung durch die logarithmische Funktion waren die Trennschärfen besonders für die mittelschwierigen und schwierigen Items höher, wenn die Items drei Antwortoptionen hatten, ebenso in den Bedingungen mit Multiple-Choice.

11.3.4 Reliabilität

Zur Bestimmung der internen Konsistenz des Englischtests wurde der Cronbach-α-Koeffizient (Cronbach, 1951) als Maß der Reliabilität für die Punktauszahlung je Item ermittelt. Die Prüfungen auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten erfolgten mithilfe des Programms *alphatst.exe* (Lautenschlager, 1989) nach der Methode von Feldt, Woodruff und Salih (1987). Tabelle 11-10 enthält die Cronbach-α-Koeffizienten für die Versuchsbedingungen, auch für standardisierte Items. Tabelle 11-11 zeigt die α-Koeffizienten nach Größe absteigend sortiert. Die Darstellung der α-Koeffizienten in verschiedenen Spalten zeigt dabei, welche Koeffizienten auf einem Niveau von 5% signifikant verschieden waren.

Tabelle 11-10

Cronbach-α-Koeffizienten, ermittelt über alle 36 Items des Tests.

Bedingung	α	α für standardisierte Items	N
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	0,82	0,82	1100
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	0,87	0,87	1100
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	0,78	0,78	1100
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	0,82	0,82	1100
MC-2 Antwortoptionen	0,77	0,78	1100
MC-3 Antwortoptionen	0,82	0,83	1100

Tabelle 11-11

α-Koeffizienten nach Größe absteigend sortiert. Die Darstellung der α-Koeffizienten in verschiedenen Spalten zeigt, welche Koeffizienten auf einem Niveau von 5% signifikant verschieden waren.

Bedingung	α	α	α
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	0,87		
ME-linear 0 bis 100 Punkte-2 Antwortoptionen		0,82	
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen		0,82	
MC-3 Antwortoptionen		0,82	
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen			0,78
MC-2 Antwortoptionen			0,77

Zuerst wurden die Daten in Bezug auf die Hypothese, dass sich in den Bedingungen mit Multiple-Choice eine signifikante Verringerung der Reliabilität bei der Reduzierung der Anzahl der Antwortoptionen von drei auf zwei zeigt, für den gesamten Test untersucht. Der Cronbach-α-Koeffizient der Versuchsbedingung mit Multiple-Choice und Items mit drei Antwortoptionen betrug 0,82. Durch die Reduzierung der Anzahl der Antwortoptionen von drei auf zwei verringerte sich α signifikant um 0,05 auf 0,77 (X^2=155,719; *df*=1; *p*<,00025).

Entgegen der Erwartung verringerte sich in beiden Auswertebedingungen des Tests mit Multipler Evaluation die interne Konsistenz ebenfalls signifikant für die Items mit zwei Antwortoptionen, im Vergleich zu denen mit drei Antwortoptionen. Bei der Auswertung durch die lineare Funktion verringerte sie sich signifikant um 0,05, von α=0,87 für die Items mit drei Antwortoptionen auf α=0,82 für die Items mit zwei Antwortoptionen (X^2=27,373; *df*=1; *p*=,00001). Wurden die Punkte anhand der logarithmischen Funktion ausgezahlt, so verringerte sich die Reliabilität ebenfalls signifikant um 0,04 von α=0,82 auf α=0,78, wenn die Anzahl der Antwortoptionen verringert wurde (X^2=10,440; *df*=1; *p*=,00164).

Der Vergleich der Multiplen Evaluation mit Multiple-Choice zeigte mit der Auswertung durch die logarithmische Funktion keine signifikante Verbesserung von Cronbach-α. Tabelle 11-12 zeigt die Ergebnisse der Prüfungen auf Signifikanz.

Tabelle 11-12

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten des Tests mit Multipler Evaluation im Vergleich zu dem Test mit Multiple-Choice.

Anzahl der Antwortoptionen	Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
2	MC	0,77	ME-linear 0 bis 100 Punkte	0,82	**0,05****	15,564	1	,00025
			ME-logarithmisch -300 bis 100 Punkte	0,78	**0,01**	0,513	1	,48085
3	MC	0,82	ME-linear 0 bis 100 Punkte	0,87	**0,05****	27,373	1	,00001
			ME-logarithmisch -300 bis 100 Punkte	0,82	**0**	0	1	,30455

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, **=signifikant auf einem Signifikanzniveau von 1%.

Wurden die Punkte durch die lineare Funktion ausgezahlt, so war die Reliabilität sowohl für die Items mit zwei, als auch für die mit drei Antwortoptionen signifikant um 0,05 höher als in den Bedingungen mit Multiple-Choice.

Die Reliabilität des Tests mit Multipler Evaluation mit Items mit zwei Antwortoptionen betrug bei einer linearen Auswertung 0,82 und mit einer logarithmischen war sie mit α=0,78 um 0,04 signifikant geringer. Hatten die Items drei Antwortoptionen, so betrug α bei einer linearen Auswertung 0,87 und war damit um 0,05 signifikant höher als der α-Koeffizient bei einer logarithmischen Auswertung. Tabelle 11-13 zeigt die Ergebnisse der Signifikanzprüfungen. Die Reliabilität des Tests mit Multipler Evaluation war also bei einer logarithmischen Auswertung signifikant geringer als bei einer linearen.

Tabelle 11-13

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten des Tests mit Multipler Evaluation mit einer linearen im Vergleich zu einer logarithmischen Auswertung.

Anzahl der Antwortoptionen	Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
2	ME-linear 0 bis 100 Punkte	0,82	ME-logarithmisch -300 bis 100 Punkte	0,78	**-0,04****	10,45	1	0,00164
3		0,87		0,82	**-0,05****	27,39	1	0,00001

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, **=signifikant auf einem Signifikanzniveau von 1%.

Im Folgenden wird die Reliabilität für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items getrennt betrachtet. Tabelle 11-14 zeigt die Cronbach-α-Koeffizienten, auch für standardisierte Items. Zuerst wurde der Einfluss der Anzahl der Antwortoptionen je Item auf die interne Konsistenz des Tests geprüft. In den Bedingungen mit Multiple-Choice zeigte sich eine signifikante Verringerung der Reliabilität aufgrund der Reduzierung der Anzahl der Antwortoptionen von drei auf zwei für einfache und mittelschwierige Items. Dabei sank α für einfache Items von 0,67 auf 0,60 ($\Delta\alpha$=-0,07; X^2=8,580; *df*=1; *p*=,00375). Für die mittelschwierigen Items zeigte sich die deutlichste Verringerung der Reliabilität um 0,17 von einem α=0,69 auf α=0,52 (X^2=44,006; *df*=1; *p*<,00001). Die Verringerung der Reliabilität der schwierigen Items von α=0,60 auf α=0,58 war nicht signifikant (X^2=0,553; *df*=1; *p*=,46395).

Tabelle 11-14

Cronbach-α-Koeffizienten der Punktauszahlung je Item für einfache, mittelschwierige und schwierige Items.

	α			α für standardisierte Items			
Schwierigkeit	einfach	mittel	schwierig	einfach	mittel	schwierig	N
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	0,68	0,61	0,68	0,68	0,61	0,67	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	0,73	0,76	0,73	0,74	0,76	0,73	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	0,64	0,54	0,62	0,66	0,54	0,62	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	0,67	0,64	0,67	0,69	0,64	0,67	1100
MC-2 Antwortoptionen	0,60	0,52	0,58	0,61	0,52	0,58	1100
MC-3 Antwortoptionen	0,67	0,69	0,60	0,68	0,69	0,60	1100

Tabelle 11-15 zeigt die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den α-Koeffizienten für die Bedingungen mit Multipler Evaluation und Items mit zwei und drei Antwortoptionen. Bei der getrennten Betrachtung der einfachen, mittelschwierigen und schwierigen Items mit einer linearen Auswertung

Tabelle 11-15

Prüfung auf signifikante Unterschiede der Cronbach-α-Koeffizienten der Bedingungen mit Multipler Evaluation mit der linearen und der logarithmischen Auswertung für Items mit zwei und drei Antwortoptionen.

Schwierig-keit	Auswertung	Bedingung 1	α 1	Bedingung 2	α 2	**$\Delta\,\alpha$**	X^2	*df*	*p*
einfach	ME-linear 0 bis 100 Punkte	2 Antwort-optionen	0,68	3 Antwort-optionen	0,73	**0,05****	6,70	1	,00947
mittel			0,61		0,76	**0,15****	54,15	1	<,00001
schwierig			0,68		0,73	**0,05****	6,70	1	,00947
einfach	ME-logarithmisch -300 bis 100 Punkte	2 Antwort-optionen	0,64	3 Antwort-optionen	0,67	**0,03**	1,76	1	,18153
mittel			0,54		0,64	**0,10****	13,92	1	,00043
schwierig			0,62		0,67	**0,05***	4,62	1	,02964

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\,\alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung. *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

fand sich bei allen Schwierigkeitsstufen eine signifikante Verringerung der Reliabilität, wenn die Anzahl der Antwortoptionen von drei auf zwei reduziert wurde. Wurden die Punkte durch die logarithmische Funktion ausgezahlt, war die Verringerung der Reliabilität für mittelschwierige und schwierige Items signifikant. Die höchste jeweils signifikante Verringerung zeigte sich in beiden Auswertebedingungen bei den mittelschwierigen Items.

Tabelle 11-16 zeigt die Einzelvergleiche der α-Koeffizienten der Bedingungen mit Multipler Evaluation und der mit Multiple-Choice. Diese Vergleiche zeigten für die Items mit zwei Antwortoptionen, dass zwar beide Auswertebedingungen des Tests mit Multipler Evaluation im Vergleich zu Multiple-Choice eine verbesserte Reliabilität für einfache, mittelschwierige und schwierige Items bewirkten, jedoch waren diese Verbesserungen nur für die Bedingungen mit der Auswertung durch die lineare Funktion signifikant. Bei den Items mit drei Antwortoptionen war die interne Konsistenz bei einer Auswertung anhand der linearen Funktion ebenfalls für

Tabelle 11-16

Prüfung auf signifikante Unterschiede der Cronbach-α-Koeffizienten der Bedingungen mit Multiple-Choice und mit Multipler Evaluation mit der linearen und der logarithmischen Auswertung für Items mit zwei und drei Antwortoptionen.

Anzahl der Antwort-optionen	Bedingung 1	Schwierig-keit	α 1	Bedingung 2	α 2	**$\Delta\,\alpha$**	X^2	*df*	*p*
2	MC	einfach	0,60	ME-linear 0 bis 100 Punkte	0,68	**0,08****	11,54	1	,00105
		mittel	0,52		0,61	**0,09****	10,00	1	,00199
		schwierig	0,58		0,68	**0,10****	17,12	1	,00015
		einfach	0,60	ME-logarithmisch -300 bis 100 Punkte	0,64	**0,04**	2,58	1	,10428
		mittel	0,52		0,54	**0,02**	0,42	1	,52364
		schwierig	0,58		0,62	**0,04**	2,36	1	,12314
3	MC	einfach	0,67	ME-linear 0 bis 100 Punkte	0,73	**0,06****	9,34	1	,00265
		mittel	0,69		0,76	**0,07****	15,17	1	,00028
		schwierig	0,60		0,73	**0,13****	35,62	1	<,00001
		einfach	0,67	ME-logarithmisch -300 bis 100 Punkte	0,67	**0**	-		-
		mittel	0,69		0,64	**-0,05***	5,19	1	,02140
		schwierig	0,60		0,67	**0,07****	8,58	1	,00375

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\,\alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

einfache, mittelschwierige und schwierige Items signifikant höher als in den Bedingungen mit Multiple-Choice. Wurden die Punkte durch die logarithmische Funktion ausgezahlt, so zeigte sich keine Verbesserung der Reliabilität für einfache Items. Für die mittelschwierigen war die interne Konsistenz signifikant verringert. Nur bei schwierigen Items zeigte sich eine signifikant höhere Reliabilität als in der Bedingung mit Multiple-Choice.

Nachfolgend wird die Reliabilität der einfachen, mittelschwierigen und schwierigen Items der Bedingungen mit Multipler Evaluation mit einer logarithmischen Auswertung im Vergleich zu einer linearen Auswertung betrachtet. Alle Bedingungen mit der linearen Auswertung zeigten dabei eine signifikant höhere Reliabilität als die Bedingungen mit der logarithmischen Auswertung, sowohl für Items mit zwei als auch für die mit drei Antwortoptionen. Eine Ausnahme war dabei nur die Reliabilität der einfachen Items mit zwei Antwortoptionen. In diesen Bedingungen war die Reliabilität zwar auch unter der linearen Auswertung im Vergleich zur logarithmischen verbessert, jedoch war diese Verbesserung nicht signifikant. Tabelle 11-17 zeigt den Vergleich der Cronbach-α-Koeffizienten. Die größte Verbesserung der Reliabilität um 0,12 zeigte sich für die schwierigen Items mit drei Antwortoptionen.

Tabelle 11-17

Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten des Tests mit Multipler Evaluation mit einer linearen im Vergleich zu einer logarithmischen Auswertung für einfache, mittelschwierige und schwierige Items.

Anzahl der Antwort-optionen	Bedingung 1	Schwierig-keit	α 1	Bedingung 2	α 2	**$\Delta\ \alpha$**	X^2	*df*	*p*
2	ME-linear 0 bis 100 Punkte	einfach	0,68	ME-logarithmisch -300 bis 100 Punkte	0,64	**-0,04**	3,22	1	0,06892
3		einfach	0,73		0,67	**-0,06****	9,34	1	0,00264
2		mittel	0,61		0,54	**-0,07***	6,33	1	0,01149
3		mittel	0,76		0,64	**-0,12****	37,93	1	<0,00001
2		schwierig	0,68		0,62	**-0,06****	6,86	1	0,00872
3		schwierig	0,73		0,67	**-0,06****	9,34	1	0,00264

α 1=Cronbach-α der ersten Bedingung des Einzelvergleichs, α 2=Cronbach-α der zweiten Bedingung, $\Delta\ \alpha$=Differenz zwischen Bedingung 1 und 2 (α 2- α 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für α in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Zusammenfassend wurde für den Test mit Multiple-Choice bestätigt, dass die Reliabilität signifikant abnimmt, wenn die Anzahl der Antwortoptionen von drei auf zwei reduziert wird. Entgegen der Erwartung verringerte sich die Reliabilität in den Bedingungen mit Multipler Evaluation ebenfalls signifikant aufgrund der Verringerung der Anzahl der Antwortoptionen, und zwar sowohl mit der linearen als auch mit einer logarithmischen Auswertung. Mit einer linearen Auswertung zeigte der Test mit Multipler Evaluation eine signifikant höhere Reliabilität als mit einer logarithmischen und zwar sowohl für die Items mit drei als auch für die mit zwei Antwortoptionen.

11.3.5 Validität

Auch im dritten Experiment schätzten die Teilnehmer ihre Fähigkeiten in der englischen Sprache auf einer Skala von eins bis sechs in den Kategorien „englische Sprache lesen“, „englische Sprache schreiben“, „englische Sprache sprechen“ und „englische Sprache verstehen“ selbst ein (siehe Abschnitt 10.2.3). Wie Tabelle 11-18 zeigt, schätzten die Testteilnehmer ihre Fähigkeit, die englische Sprache lesen zu können, mit einem Mittelwert von 4,19 am höchsten ein. Die englische Sprache verstehen zu können beurteilten die Teilnehmer mit einem durchschnittlichen Wert von 3,97 am zweithöchsten. In der Kategorie „englische Sprache sprechen“ schätzten die Teilnehmer ihre Fähigkeiten durchschnittlich mit 3,60 ein. In der Kategorie „englische Sprache schreiben“ schätzten sie ihre Fähigkeit mit einem Mittelwert von 3,33 am geringsten ein.

Tabelle 11-18

Deskriptive Statistik der Selbsteinschätzung der Testteilnehmer, ermittelt über alle Bedingungen des Experiments.

Kategorie Selbsteinschätzung	Minimum	Maximum	Mittelwert	Standard-abweichung	Standard-fehler	Median	N
Lesen	1	6	**4,19**	1,26	0,015	4,0	6600
Schreiben	1	6	**3,33**	1,18	0,014	3,0	6600
Sprechen	1	6	**3,60**	1,26	0,015	4,0	6600
Verstehen	1	6	**3,97**	1,18	0,014	4,0	6600

Über alle vier Kategorien der selbsteingeschätzten Fähigkeiten wurde jeweils eine zweifaktorielle Varianzanalyse berechnet. Dabei waren die Faktoren die Anzahl der Antwortoptionen in den Stufen zwei und drei Antwortoptionen und die verwendete Auswertefunktion in den drei Stufen MC-dichotome Auswertung (0 oder 1 Punkt), ME-lineare Auswertung (0 bis 100 Punkte) und ME-logarithmische Auswertung (-300 bis 100 Punkte). Zwischen allen sechs Versuchsbedingungen des Experiments zeigten sich dabei keine signifikanten Unterschiede in Bezug auf die selbsteingeschätzten Englischkenntnisse der Teilnehmer. Eine Ausnahme war eine signifikante Interaktion zwischen der verwendeten Auswertefunktion und der Anzahl der Antwortoptionen in der Kategorie „englische Sprache verstehen“. Tabelle 11-19 zeigt die Ergebnisse von vier zweifaktoriellen Varianzanalysen jeweils über die selbsteingeschätzten Fähigkeiten mit den Faktoren „Auswertefunktion“ und „Anzahl der Antwortoptionen“ für die vier Kategorien der selbsteingeschätzten Fähigkeiten.

Als Maß der Validität wurde die Korrelation (nach Pearson) zwischen den selbsteingeschätzten Englischfähigkeiten und den Punktsummen der Teilnehmer bestimmt. Dazu wurden die Werte der selbsteingeschätzten Fähigkeiten zunächst

Tabelle 11-19

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Bedingungen mithilfe einer zweifaktoriellen Varianzanalyse.

Kategorie Selbsteinschätzung	Faktor	F(1, 6599)	p	η^2
Lesen	Auswertefunktion	0,747	0,474	<0,001
Lesen	Anzahl der Antwortoptionen	0,359	0,549	<0,001
Lesen	Auswertefunktion * Anzahl der Antwortoptionen	1,208	0,299	<0,001
Schreiben	Auswertefunktion	0,224	0,799	<0,001
Schreiben	Anzahl der Antwortoptionen	0,297	0,586	<0,001
Schreiben	Auswertefunktion * Anzahl der Antwortoptionen	1,661	0,190	<0,001
Sprechen	Auswertefunktion	0,011	0,989	<0,001
Sprechen	Anzahl der Antwortoptionen	0,075	0,784	<0,001
Sprechen	Auswertefunktion * Anzahl der Antwortoptionen	0,900	0,407	<0,001
Verstehen	Auswertefunktion	0,883	0,414	<0,001
Verstehen	Anzahl der Antwortoptionen	0,021	0,884	<0,001
Verstehen	Auswertefunktion * Anzahl der Antwortoptionen	3,053	0,047*	0,001

*=signifikant auf einem Signifikanzniveau von 5%.

z-transformiert und der Mittelwert (z-Index) über alle vier Kategorien gebildet. Dieser z-Index wurde mit den Punktsummen der Testteilnehmer korreliert. Um die Korrelationskoeffizienten der Bedingungen auf signifikante Unterschiede zu prüfen, wurde die Methode nach Bortz (2005, S. 220 f.) angewendet. Dazu wurden die Korrelationskoeffizienten Fisher-Z-transformiert. Tabelle 11-20 zeigt die Korrelationskoeffizienten und die Fisher-Z-transformierten Koeffizienten der selbsteingeschätzten Fähigkeiten mit den Punktsummen der 36 Items des Tests. Den höchsten Korrelationskoeffizienten zeigte mit r=,64 (Z(r)=0,75) die Bedingung mit Multipler Evaluation und Items mit drei Antwortoptionen sowie der Auszahlung der Punkte durch die lineare Funktion. Den niedrigsten Korrelationskoeffizienten r=,52 (Z(r)=0,58) hatte die Bedingung mit Multipler Evaluation, Items mit drei

Tabelle 11-20

Koeffizienten nach Pearson und Fisher-Z-transformierte Korrelationskoeffizienten des Mittelwerts (z-Index) der selbsteingeschätzten Fähigkeiten der Teilnehmer mit den Punktsummen aller 36 Items des Tests.

Bedingung	r	Z(r)	N
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	,61	0,70	1100
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	,64	0,75	1100
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	,55	0,61	1100
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	,52	0,58	1100
MC-2 Antwortoptionen	,61	0,71	1100
MC-3 Antwortoptionen	,62	0,73	1100

Tabelle 11-21

Fisher-Z-transformierte Korrelationskoeffizienten nach Größe absteigend sortiert. Die Darstellung der Koeffizienten in verschiedenen Spalten zeigt, welche Koeffizienten auf einem Niveau von 5% signifikant voneinander verschieden waren.

Bedingung	r	Z(r)	r	Z(r)
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	,64	0,75		
MC-3 Antwortoptionen	,62	0,73		
MC-2 Antwortoptionen	,61	0,71		
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	,61	0,70		
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen			,55	0,61
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen			,52	0,58

Antwortoptionen und der Auswertung mithilfe der logarithmischen Funktion. Tabelle 11-21 zeigt die Korrelationskoeffizienten nach Größe absteigend sortiert. Die Anordnung in verschiedenen Spalten bedeutet, dass die Koeffizienten auf einem Niveau von 5% signifikant verschieden waren.

Zuerst soll für den gesamten Test geklärt werden, ob sich durch die Reduzierung der Anzahl der Antwortoptionen von drei auf zwei die Validität in den verschiedenen Auswertebedingungen signifikant verringerte. Für den Test mit Multiple-Choice und Items mit drei Antwortoptionen betrug Z(r)=0,73 (r=,62). In der Bedingung mit Multiple-Choice und Items mit zwei Antwortoptionen war Z(r) um 0,02 geringer. Diese Verschlechterung des Z(r) war nicht signifikant (*z*=0,496; *df*=1; n.*s.*). Für die Bedingung mit Multipler Evaluation, Items mit drei Antwortoptionen und der Auswertung durch die lineare Funktion war Z(r) mit 0,75 (r=,64) um 0,05 höher, jedoch nicht signifikant, als Z(r) mit 0,70 (r=,61) für die Items mit zwei Antwortoptionen (*z*=1,171; *df*=1; n.s.). Die Bedingung mit der Auswertung durch die logarithmische Funktion zeigte für Items mit drei Antwortoptionen und Z(r)=0,58 (r=,52) dagegen ein um 0,03, nicht signifikant geringeres Z(r) als die Bedingung mit den Items mit zwei Antwortoptionen und einem Z(r)=0,61 (r=,55) (*z*=-0,703; *df*=1; n.s.). Im Vergleich zu den Bedingungen mit Multiple-Choice war die Validität der Multiplen Evaluation mit der Auswertung durch die lineare Funktion geringfügig,

Tabelle 11-22

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen mit Multiple-Choice und Multipler Evaluation (nach Bortz, 2005) für alle 36 Items.

Anzahl der Antwortoptionen	Bedingung 1	Z(r) 1	Bedingung 2	Z(r) 2	**Δ Z(r)**	z	*df*	*p*
2	MC	0,71	ME-linear 0 bis 100 Punkte	0,70	**0,02**	0,492	1	n.s.
3		0,73		0,75	**0,02**	0,463	1	n.s.
2		0,71	ME-logarithmisch -300 bis 100 Punkte	0,61	**-0,10***	-2,292	1	0,05
3		0,73		0,58	**-0,15****	-3,534	1	0,0005

Z(r) 1=Fisher-Z-transformierte Korrelationskoeffizienten der ersten Bedingung des Einzelvergleichs, Z(r) 2 = Fisher-Z-transformierte Korrelationskoeffizienten der zweiten Bedingung. Δ Z(r)=Differenz zwischen der Bedingung 1 und 2 (Z(r) 2 - Z(r) 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung. *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

sowohl für die Items mit zwei, als auch für die mit drei Antwortoptionen jeweils um 0,02 nicht signifikant verbessert. Wurden die Punkte durch die logarithmische Funktion ausgezahlt, so war die Validität in beiden Bedingungen signifikant gegenüber Multiple-Choice verschlechtert. Für die Items mit zwei Antwortoptionen betrug diese Verschlechterung 0,10. Für die Items mit drei Antwortoptionen betrug sie 0,15. Eine signifikante Verbesserung der Validität der Multiplen Evaluation gegenüber Multiple-Choice zeigte sich also weder für die lineare noch für die logarithmische Auswertung. Die Ergebnisse der Vergleiche der Korrelationskoeffizienten der einzelnen Bedingungen zeigt Tabelle 11-22.

Der Vergleich der Validität des Tests mit Multipler Evaluation mit einer linearen Auswertung zu der mit einer logarithmischen ist in Tabelle 11-23 dargestellt. Sowohl bei den Items mit zwei als auch bei denen mit drei Antwortoptionen war die Validität des Tests signifikant verbessert, wenn die Auswertung anhand einer linearen Funktion erfolgte im Vergleich zu einer logarithmischen.

Tabelle 11-23

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen mit Multipler Evaluation und einer linearen sowie einer logarithmischen Auswertung (nach Bortz, 2005) für alle 36 Items.

Anzahl der Antwortoptionen	Bedingung 1	Z(r) 1	Bedingung 2	Z(r) 2	**Δ Z(r)**	z	*df*	*p*
2	ME-linear 0 bis 100 Punkte	0,70	ME-logarithmisch -300 bis 100 Punkte	0,61	**-0,09***	-2,11	1	0,05
3		0,75		0,58	**-0,17****	-3,98	1	<0,0005

Z(r) 1=Fisher-Z-transformierte Korrelationskoeffizienten der ersten Bedingung des Einzelvergleichs, Z(r) 2 = Fisher-Z-transformierte Korrelationskoeffizienten der zweiten Bedingung. Δ Z(r)=Differenz zwischen der Bedingung 1 und 2 (Z(r) 2 - Z(r) 1). Positive Differenzen bedeutetn eine Verbesserung des Wertes für Z(r) in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung. *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%, k=Anzahl der Antwortoptionen.

Im Folgenden wird die Validität getrennt für die drei Schwierigkeitsstufen der Items betrachtet. Tabelle 11-24 zeigt die Koeffizienten nach Pearson und die Fisher-Z-transformierten Koeffizienten des z-Index der selbsteingeschätzten Fähigkeiten mit den Punktsummen der einfachen, mittelschwierigen und schwierigen Items. Der Vergleich der Bedingungen mit Items mit zwei und der Bedingungen mit Items mit drei Antwortoptionen zeigte für die Bedingungen mit Multiple-Choice für einfache

Tabelle 11-24

Korrelationskoeffizienten nach Pearson und Fisher-Z-transformierte Korrelationskoeffizienten des z-Index der selbsteingeschätzten Fähigkeiten mit den Punktsummen der einfachen, mittelschwierigen und schwierigen Items.

Schwierigkeit	einfach (A)		mittel (B)		schwierig (C)		
Bedingung	r	Z(r)	r	Z(r)	r	Z(r)	N
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	,54	0,61	,47	0,52	,49	0,54	1100
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	,55	0,61	,56	0,64	,50	0,55	1100
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	,51	0,56	,43	0,46	,37	0,39	1100
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	,53	0,59	,48	0,52	,29	0,30	1100
MC-2 Antwortoptionen	,55	0,62	,46	0,50	,50	0,55	1100
MC-3 Antwortoptionen	,52	0,58	,57	0,64	,47	0,50	1100

Tabelle 11-25

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen (nach Bortz, 2005) mit Items mit zwei und drei Antwortoptionen für einfache, mittelschwierige und schwierige Items.

Schwierig-keit	Bedingung 1	Z(r) 1	Bedingung 2	Z(r) 2	**Δ Z(r)**	z	*df*	*p*
einfach	MC 2 Antwort-optionen	0,62	MC 3 Antwort-optionen	0,58	**-0,04**	-0,951	1	n.s.
mittel		0,50		0,64	**0,14****	3,286	1	<,0005
schwierig		0,55		0,50	**-0,04**	-1,006	1	n.s.
einfach	ME-linear 0 bis 100 Punkte 2 Antwortoptionen	0,61	ME-linear 0 bis 100 Punkte 3 Antwortoptionen	0,61	**0**	0	1	n.s.
mittel		0,52		0,64	**0,12****	2,892	1	0,005
schwierig		0,54		0,55	**0,01**	0,234	1	n.s.
einfach	ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	0,56	ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	0,59	**0,03**	0,703	1	n.s.
mittel		0,46		0,52	**0,06**	1,405	1	n.s.
schwierig		0,39		0,30	**-0,09***	-2,108	1	0,05

Z(r) 1=Fisher-Z-transformierte Korrelationskoeffizienten der ersten Bedingung des Einzelvergleichs, Z(r) 2=Fisher-Z-transformierte Korrelationskoeffizienten der zweiten Bedingung, Δ Z(r)=Differenz zwischen den Bedingungen 1 und 2 (Z(r) 2- Z(r) 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung. *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

und schwierige Items keine signifikanten Unterschiede zwischen den Bedingungen. Für die mittelschwierigen Items mit drei Antwortoptionen war Z(r) mit 0,64 (r=,56) um 0,14 signifikant höher als Z(r)=0,50 (r=,46) bei Items mit zwei Antwortoptionen.

Tabelle 11-25 zeigt die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen für die einfachen, mittelschwierigen und schwierigen Items. Für die Bedingungen mit Multipler Evaluation und der Auswertung durch die lineare Funktion zeigten sich ebenfalls keine signifikanten Unterschiede für einfache und schwierige Items in Abhängigkeit von der Anzahl der Antwortoptionen. Bei den mittelschwierigen Items betrug Z(r)=0,64 (r=,56) für die Items mit drei Antwortoptionen und war damit um 0,12 signifikant höher als die Validität der Bedingung mit zwei Antwortoptionen mit Z(r)=0,52 (r=,47). Wurden die Punkte anhand der logarithmischen Funktion ausgezahlt, zeigte sich kein signifikanter Unterschied zwischen den Bedingungen für einfache und mittelschwierige Items. Bei schwierigen Items verschlechterte sich die Validität signifikant um 0,09, wenn die Items drei statt zwei Antwortoptionen hatten.

Tabelle 11-26

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen mit Multipler Evaluation und Multiple-Choice für einfache, mittelschwierige und schwierige Items.

Schwierig-keit	Bedingung 1	Z(r) 1	Bedingung 2	Z(r) 2	**Δ Z(r)**	z	*df*	*p*
einfach	MC-2 Antwortoptionen	0,62	ME-linear 0 bis 100 Punkte 2 Antwortoptionen	0,61	**-0,01**	-0,233	1	n.s.
mittel		0,50		0,52	**0,02**	0,360	1	n.s.
schwierig		0,55		0,54	**-0,01**	-0,124	1	n.s.
einfach	MC-3 Antwortoptionen	0,58	ME-linear 0 bis 100 Punkte 3 Antwortoptionen	0,61	**0,03**	0,703	1	n.s.
mittel		0,64		0,64	**0**	0	1	n.s.
schwierig		0,50		0,55	**0,05**	1,171	1	n.s.
einfach	MC-2 Antwortoptionen	0,62	ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	0,56	**-0,06**	-1,334	1	n.s.
mittel		0,50		0,46	**-0,04**	-0,933	1	n.s.
schwierig		0,55		0,39	**-0,16****	-3,669	1	<0,0005
einfach	MC-3 Antwortoptionen	0,58	ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	0,59	**0,01**	0,234	1	n.s.
mittel		0,64		0,52	**-0,12****	-2,814	1	0,005
schwierig		0,50		0,30	**-0,20****	-4,684	1	<0,0005

Z(r) 1=Fisher-Z-transformierte Korrelationskoeffizienten der ersten Bedingung des Einzelvergleichs, Z(r) 2=Fisher-Z-transformierte Korrelationskoeffizienten der zweiten Bedingung, Δ Z(r)=Differenz zwischen den Bedingungen 1 und 2 (Z(r) 2- Z(r) 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung. **=signifikant auf einem Signifikanzniveau von 1%.

Wurden die Korrelationskoeffizienten der Bedingungen mit Multipler Evaluation mit denen der Bedingungen mit Multiple-Choice verglichen, so zeigten sich keine signifikanten Unterschiede zwischen den Bedingungen für einfache, mittelschwierige und schwierige Items bei einer linearen Auswertung. Tabelle 11-26 zeigt die Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen mit Multipler Evaluation und Multiple-Choice für einfache, mittelschwierige und schwierige Items. Signifikante Unterschiede zwischen den Bedingungen mit Multiple-Choice und mit Multipler Evaluation zeigten sich bei der Auswertung durch die logarithmische Funktion. Bei den Items mit zwei Antwortoptionen war die Validität des Tests mit Multipler Evaluation bei den schwierigen Items signifikant um 0,16 geringer als die des Tests mit Multiple-Choice. Bei den mittelschwierigen Items mit drei Antwortoptionen war Z(r) mit 0,52 (r=,48) in der Bedingung mit Multipler Evaluation signifikant um 0,12 verringert, gegenüber der mit Multiple-Choice mit Z(r)=0,64 (r=,56). Für schwierige Items mit drei Antwortoptionen war die Validität der Bedingungen mit Multipler Evaluation ebenfalls signifikant um 0,20 geringer als die der Bedingung mit Multiple-Choice.

Tabelle 11-27

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Korrelationskoeffizienten der Bedingungen mit Multipler Evaluation und einer linearen und logarithmischen Auswertung für einfache, mittelschwierige und schwierige Items.

Schwierig-keit	Anzahl der Antwort-optionen	Bedingung 1	Z(r) 1	Bedingung 2	Z(r) 2	**Δ Z(r)**	z	*df*	*p*
einfach	2	ME-linear 0 bis 100 Punkte	0,61	ME-logarithmisch -300 bis 100 Punkte	0,56	**-0,05**	-1,19	0	n.s.
	3		0,61		0,59	**-0,03**	-0,66	1	n.s.
mittel	2		0,52		0,46	**-0,05**	-1,21	1	n.s.
	3		0,64		0,52	**-0,12****	-2,74	1	0,005
schwierig	2		0,54		0,39	**-0,15****	-3,55	1	<0,0005
	3		0,55		0,30	**-0,24****	-5,68	2	<0,0005

Z(r) 1=Fisher-Z-transformierte Korrelationskoeffizienten der ersten Bedingung des Einzelvergleichs, Z(r) 2=Fisher-Z-transformierte Korrelationskoeffizienten der zweiten Bedingung, Δ Z(r)=Differenz zwischen den Bedingungen 1 und 2 (Z(r) 2- Z(r) 1). Positive Differenzen bedeuten eine Verbesserung des Wertes für Z(r) in der Bedingung 2 gegenüber der Bedingung 1, negative Differenzen eine Verschlechterung. **=signifikant auf einem Signifikanzniveau von 1%.

Der Vergleich der Bedingungen mit Multipler Evaluation und einer linearen Auswertung mit einer logarithmischen zeigte auch bei der getrennten Betrachtung der einfachen, mittelschwierigen und schwierigen Items eine höhere Validität bei einer linearen Auswertung. Signifikant war diese Verbesserung bei den mittelschwierigen Items mit drei Antwortoptionen und bei den schwierigen sowohl mit zwei als auch mit drei Antwortoptionen. Tabelle 11-27 zeigt die Vergleiche zwischen den Bedingungen mit einer linearen und einer logarithmischen Auswertung.

Zusammenfassend zeigte sich für den gesamten Test keine Verringerung der Validität aufgrund der Reduzierung der Anzahl der Antwortoptionen von zwei auf drei, weder in den Bedingungen mit Multiple-Choice noch in denen mit Multipler Evaluation. Nur wenn ausschließlich die schwierigen Items betrachtet wurden, zeigte sich in allen Bedingungen eine signifikante Verschlechterung der Validität aufgrund der Reduzierung der Anzahl der Antwortoptionen. Der Vergleich des Tests mit Multipler Evaluation mit einer linearen und einer logarithmischen Auswertung zeigte für die lineare Auswertung eine signifikant höhere Reliabilität, besonders bei schwierigen Items.

11.3.6 Realismusindex

Als Maß der Güte der Kalibrierung eines Testteilnehmers wurde der individuelle Realismusindex bestimmt. Der niedrigste gemessene Realismusindex betrug -0,45, der höchste 1,47. Tabelle 11-28 zeigt die deskriptive Statistik des Realismusindex für die Versuchsbedingungen. Abbildung 11-5 zeigt den Realismusindex für die Versuchsbedingungen unterteilt in zehn Perzentile. Zunächst wird der Realismusindex im Bezug auf die unabhängige Variable „Anzahl der Antwortoptionen“ betrachtet. Die Bedingung mit Items mit drei Antwortoptionen zeigte mit a=0,70 einen durchschnittlich um 0,04 höheren Realismusindex im Vergleich zum Realismusindex von a=0,66 der Bedingung mit Items mit zwei Antwortoptionen. In einer zweifaktoriellen Varianzanalyse über den Realismusindex mit den Faktoren „Anzahl der Antwortoptionen“ und „Auswertefunktion“ war dieser Unterschied für den Faktor „Anzahl der Antwortoptionen“ signifikant ($F(1, 4399)=1{,}616$; $p<{,}001$; $\eta^2={,}006$). Bei der Betrachtung des Realismusindex für die

Tabelle 11-28

Deskriptive Statistik des Realismusindex.

Bedingungen	Minimum	Maximum	**Mittelwert**	Standardfehler	Standardabweichung	Median	Anzahl Items	N
ME-2 Antwortoptionen	-0,45	1,47	**0,66**	0,006	0,26	0,71	36	2200
ME-3 Antwortoptionen	-0,25	1,21	**0,70**	0,005	0,23	0,75	36	2200
ME-linear 0 bis 100 Punkte	-0,35	1,21	**0,64**	0,005	0,24	0,68	36	2200
ME-logarithmisch -300 bis 100 Punkte	-0,45	1,47	**0,73**	0,005	0,24	0,78	36	2200
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	-0,35	1,12	**0,61**	0,008	0,26	0,65	36	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	-0,24	1,21	**0,67**	0,006	0,22	0,70	36	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-0,45	1,47	**0,72**	0,007	0,25	0,77	36	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-0,25	1,21	**0,74**	0,007	0,23	0,80	36	1100

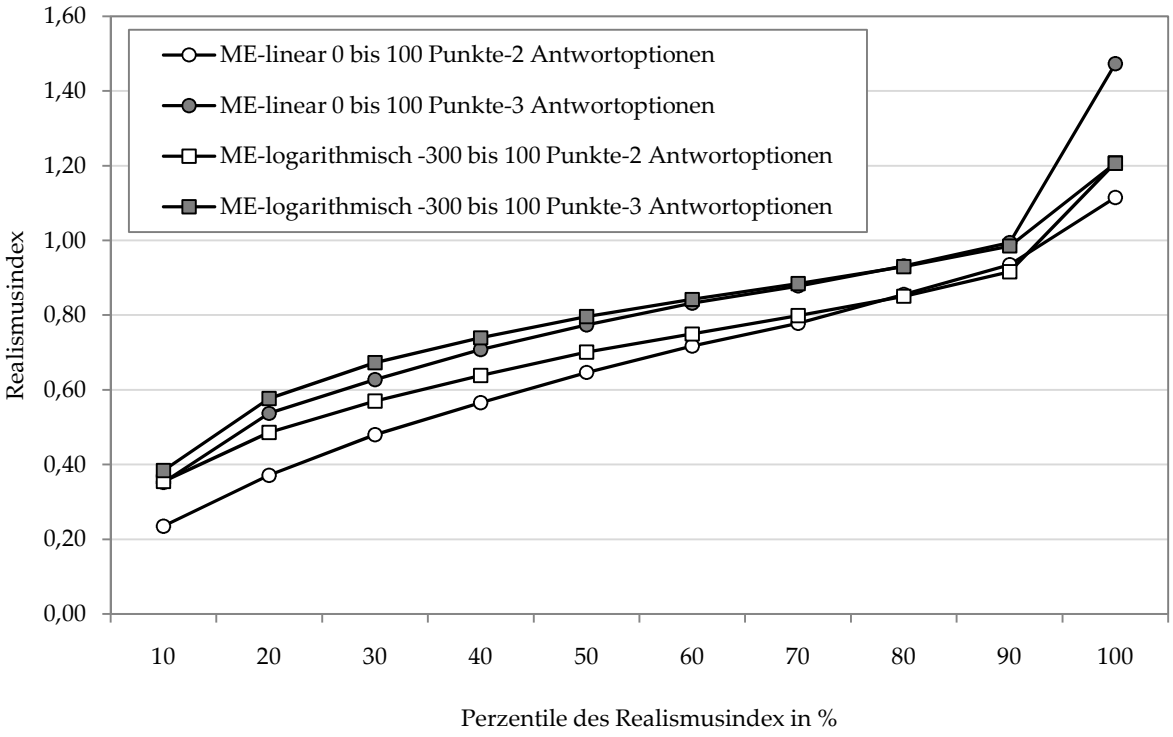

Abbildung 11-5: Realismusindex in den Versuchsbedingungen, dargestellt für zehn Perzentile.

unabhängige Variable „Auswertefunktion“ zeigten die Bedingungen mit einer logarithmischen Auswertung mit a=0,73 einen um 0,09 signifikant höheren Index als die Bedingung mit der linearen Auswertung und a=0,64. Die Signifikanz dieses Unterschieds belegte eine zweifaktorielle Varianzanalyse für den Realismusindex mit dem Faktor „Auswertefunktion“ (F(1, 4399)=172,625; p<,001; η^2=,038). Zudem fand sich eine signifikante Interaktion zwischen der Anzahl der Antwortoptionen und der verwendeten Auswertefunktion. Der Effekt war aber nur gering (F(1, 4399)=6,953; p<,008; η^2=,002). Den niedrigsten durchschnittlichen Realismusindex hatten die Teilnehmer der Bedingung „ME-linear 0 bis 100 Punkte“ mit a=0,61 und den höchsten die der Bedingung „ME-logarithmisch -300 bis 100 Punkte“ mit a=0,74.

11.3.7 Punktauszahlungen nach Realismuskorrektur

Auch im dritten Experiment wurden die Antwortsicherheiten in die richtige Antwort für jeden Testteilnehmer auf der Basis seines individuellen Realismusindex korrigiert und die Punkte auf der Basis der korrigierten Antwortsicherheiten erneut ausgezahlt. Abbildung 11-6 zeigt die korrigierten Punktsummen für den gesamten Test und je Schwierigkeitsstufe im Vergleich zu den nicht korrigierten Punktsummen. Die Tabelle mit der deskriptiven Statistik der korrigierten Punktauszahlungen befindet sich in Anhang C.3.

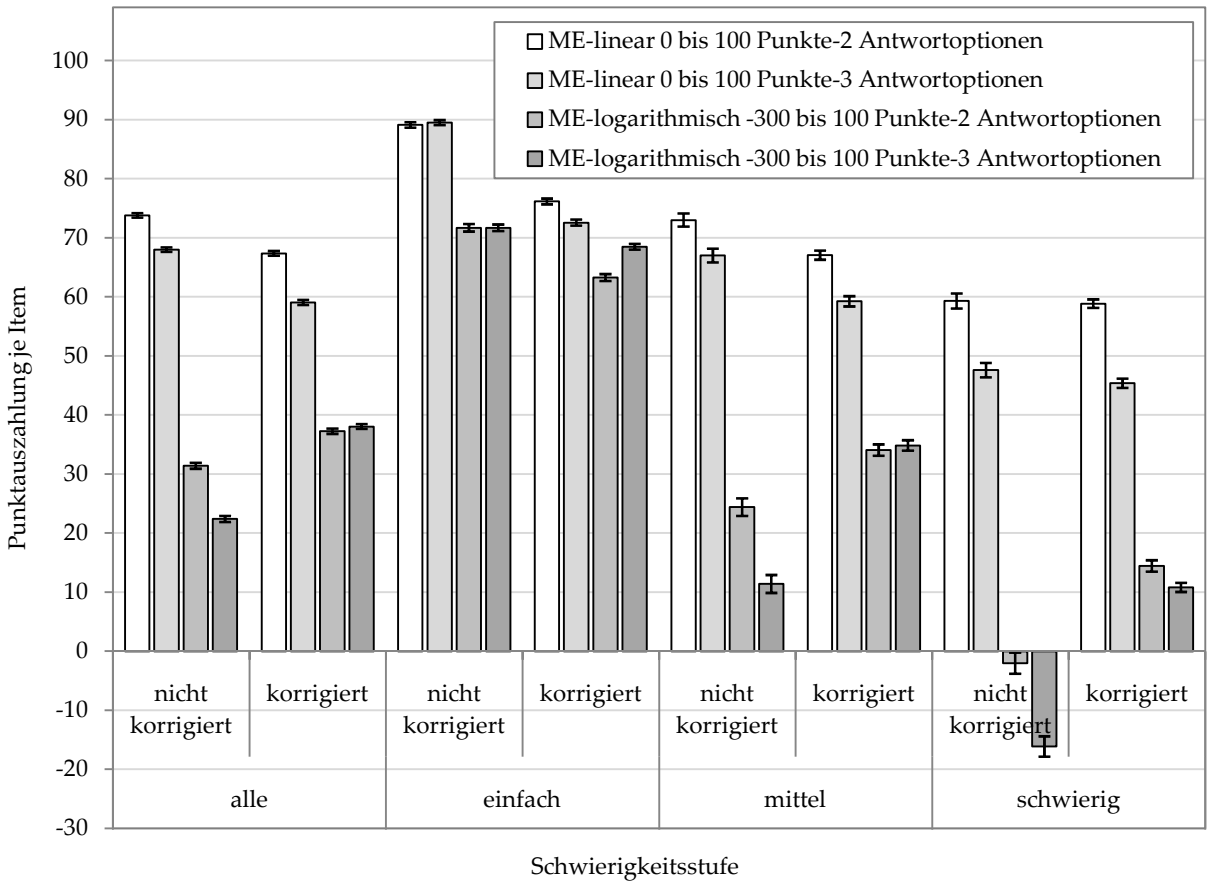

Abbildung 11-6: Mittelwert je Item der korrigierten und nicht korrigierten Punkte. Die Fehlerbalken repräsentieren die Standardfehler.

11.3.8 Reliabilität nach Realismuskorrektur

Für die anhand des Realismusindex korrigierten Punktauszahlungen je Item wurden die Cronbach-α-Koeffizienten ermittelt. Diese Koeffizienten zeigt Tabelle 11-29. Wurden alle Items des Tests betrachtet, so wurde durch die Korrektur bei einer linearen Auswertung jeweils ein Cronbach-α von 0,95 sowohl für die Bedingung mit zwei als auch drei Antwortoptionen erzielt. Damit zeigte sich eine signifikante Verbesserung der Reliabilität um 0,13 für die Bedingung mit Items mit zwei Antwortoptionen und um 0,08 für die Bedingung mit drei Antwortoptionen. Die Cronbach-α-Koeffizienten für nicht korrigierte und korrigierte Punktsummen und die Ergebnisse der Prüfungen auf signifikante Unterschiede können Tabelle 11-30 entnommen werden.

Tabelle 11-29

Cronbach-α-Koeffizienten, ermittelt für die korrigierten Punktauszahlungen.

Bedingung	Schwierigkeits-stufe	Anzahl der Antwortoptionen	α	α standardisierte Items	Anzahl Items	N
ME-linear 0 bis 100 Punkte	alle	2	0,95	0,95	36	1100
		3	0,95	0,96	36	1100
	einfach (A)	2	0,94	0,95	12	1100
		3	0,95	0,96	12	1100
	mittel (B)	2	0,84	0,85	12	1100
		3	0,89	0,89	12	1100
	schwierig (C)	2	0,81	0,81	12	1100
		3	0,83	0,83	12	1100
ME-logarithmisch -300 bis 100 Punkte	alle	2	0,87	0,89	36	1100
		3	0,86	0,89	36	1100
	einfach (A)	2	0,86	0,88	12	1100
		3	0,86	0,89	12	1100
	mittel (B)	2	0,67	0,68	12	1100
		3	0,69	0,70	12	1100
	schwierig (C)	2	0,66	0,67	12	1100
		3	0,60	0,60	12	1100

Tabelle 11-30

Ergebnisse der Prüfung auf signifikante Unterschiede zwischen den Cronbach-α-Koeffizienten der korrigierten und nicht korrigierten Punktauszahlungen.

Bedingung	Schwierig-keitsstufe	Anzahl der Antwortoptionen	α nicht korri-giert	α korri giert	$\Delta\ \alpha$	X^2	*df*	*p*
ME-linear 0 bis 100 Punkte	alle	2	0,82	0,95	**0,13****	396,056	1	<,00001
		3	0,87	0,95	**0,08****	227,525	1	<,00001
	einfach (A)	2	0,68	0,94	**0,26****	575,872	1	<,00001
		3	0,73	0,95	**0,22****	583,447	1	<,00001
	mittel (B)	2	0,61	0,84	**0,23****	177,827	1	<,00001
		3	0,76	0,89	**0,13****	137,482	1	<,00001
	schwierig (C)	2	0,68	0,81	**0,13****	62,330	1	<,00001
		3	0,73	0,83	**0,10****	49,219	1	<,00001
ME-logarithmisch -300 bis 100 Punkte	alle	2	0,78	0,87	**0,09****	70,982	1	<,00001
		3	0,82	0,86	**0,04****	16,358	1	,0002
	einfach (A)	2	0,64	0,86	**0,22****	198,945	1	<,00001
		3	0,67	0,86	**0,19****	165,130	1	<,00001
	mittel (B)	2	0,54	0,67	**0,13****	25,490	1	,00002
		3	0,64	0,69	**0,05***	5,188	1	,02140
	schwierig (C)	2	0,62	0,66	**0,04**	2,872	1	,08614
		3	0,67	0,60	**-0,07****	8,580	1	,00375

$\Delta\ \alpha$=Differenz zwischen α-korrigierte Punkte und α-nicht korrigierte Punkte (α korrigiert – α nicht korrigiert). Positive Differenzen bedeuten eine Verbesserung des korrigierten α, negative Differenzen eine Verschlechterung. *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%. N je Bedingung=1100, Anzahl Items je Schwierigkeitsstufe=12, für alle Schwierigkeitsstufen 36.

In den Bedingungen mit der Auswertung durch die logarithmische Funktion verbesserten sich die Cronbach-α-Koeffizienten ebenfalls durch die Korrektur. Für die Items mit zwei Antwortoptionen verbesserte sich α von 0,78 um 0,09 signifikant auf 0,86, für die mit drei Antwortoptionen ebenfalls signifikant um 0,04 von 0,82 auf 0,86.

Die Cronbach-α-Koeffizienten, ermittelt für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items, verbesserten sich bei einer linearen Auswertung alle signifikant aufgrund der Korrektur. In den Bedingungen mit der Auswertung mithilfe der logarithmischen Funktion waren ebenfalls alle

Cronbach-α-Koeffizienten aufgrund der Korrektur signifikant verbessert. Eine Ausnahme waren dabei aber die α-Koeffizienten der Bedingungen „ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen“ für mittelschwierige Items und die Bedingung „ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen“ für schwierige Items. Diese waren zwar ebenfalls aufgrund der Korrektur verbessert, aber nicht signifikant. Für die schwierigen Items mit drei Antwortoptionen war α bei der logarithmischen Auswertung dagegen nach der Korrektur der Punkte signifikant von 0,67 auf 0,60 um 0,07 verringert.

Zusammenfassend wurde gezeigt, dass durch die Korrektur auf der Basis des individuellen Realismusindex die Reliabilität des Tests sowohl bei einer linearen als auch bei einer logarithmischen Auswertung signifikant verbessert werden kann.

11.3.9 Validität nach Realismuskorrektur

Um die Hypothese zu prüfen, dass eine Korrektur der Antwortsicherheiten auf der Basis des Realismusindex zu einer Verbesserung der Validität führt, wurden die Korrelationen (nach Pearson) des z-Index der selbsteingeschätzten Fähigkeiten mit den Punktsummen der realismuskorrigierten Punktauszahlungen ermittelt. Für die Prüfung auf signifikante Unterschiede (Bortz, 2005) wurden die Korrelationskoeffizienten Fisher-Z-transformiert. Tabelle 11-31 zeigt die Korrelationskoeffizienten r und die Fisher-Z-transformierten Koeffizienten Z(r). Wurden die

Tabelle 11-31

Korrelationskoeffizienten nach Pearson der „Selbsteinschätzung“ mit den auf der Basis der realismuskorrigierten Punktsummen.

Schwierigkeit	alle		einfach (A)		mittel (B)		schwierig (C)		
Bedingung	r	Z(r)	r	Z(r)	r	Z(r)	r	Z(r)	N
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	,56	0,63	,54	0,61	,51	0,56	,50	0,55	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	,59	0,68	,57	0,64	,56	0,63	,50	0,55	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	,58	0,66	,55	0,62	,50	0,55	,43	0,46	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	,62	0,72	,58	0,66	,57	0,64	,41	0,44	1100

Korrelationen für die Punktsummen aller Items des Tests ermittelt, so zeigte sich eine nicht signifikante Verringerung der Validität für die Bedingungen mit der linearen Auswertung sowohl für die Items mit zwei als auch die mit drei Antwortoptionen. Bei der Auswertung mithilfe der logarithmischen Funktion war die Validität nach der Korrektur für die Items mit zwei Antwortoptionen um 0,05 höher als vor der Korrektur, jedoch nicht signifikant. Die Validität der Bedingung mit Items mit drei Antwortoptionen war signifikant um 0,14 von Z(r)=0,58 (r=,52) auf Z(r)=0,72 (r=,62) verbessert. Tabelle 11-32 zeigt die Ergebnisse der Prüfungen auf signifikante Unterschiede zwischen den auf der Basis der korrigierten und nicht korrigierten Punktsummen ermittelten Korrelationskoeffizienten.

Tabelle 11-32

Ergebnisse der Prüfung auf signifikante Unterschiede (nach Bortz, 2005) zwischen den Korrelationskoeffizienten der korrigierten und nicht korrigierten Fisher-Z-transformierten Korrelationskoeffizienten.

Bedingung	Schwierig-keitsstufe	Anzahl der Antwortoptionen	Z(r) nicht korri-giert	Z(r) korri-giert	**Δ Zr**	X^2	*df*	*p*
ME-linear 0 bis 100 Punkte	alle	2	0,70	0,63	**-0,07**	-1,634	1	n.s.
		3	0,75	0,68	**-0,07**	-1,660	1	n.s.
	einfach (A)	2	0,61	0,61	**0**	-0,031	1	n.s.
		3	0,61	0,64	**0,03**	0,671	1	n.s.
	mittel (B)	2	0,52	0,56	**0,04**	1,049	1	n.s.
		3	0,64	0,63	**-0,01**	-0,234	1	n.s.
	schwierig (C)	2	0,54	0,55	**0,01**	0,122	1	n.s.
		3	0,55	0,55	**0**	0	1	n.s.
ME-logarithmisch -300 bis 100 Punkte	alle	2	0,61	0,66	**0,05**	1,066	1	n.s.
		3	0,58	0,72	**0,14****	3,334	1	<,0005
	einfach (A)	2	0,56	0,62	**0,06**	1,405	1	n.s.
		3	0,59	0,66	**0,07**	1,639	1	n.s.
	mittel (B)	2	0,46	0,55	**0,09***	2,108	1	,05.
		3	0,52	0,64	**0,12****	2,810	1	,005
	schwierig (C)	2	0,39	0,46	**0,07***	1,709	1	,05
		3	0,30	0,44	**0,14****	3,223	1	<,0005

Z(r)=Fisher-Z-transformierte Korrelationskoeffizienten, Δ Z(r)=Differenz zwischen Z(r) korrigiert und Z(r) nicht korrigiert, *=signifikant auf einem Signifikanzniveau von 5%, **=signifikant auf einem Signifikanzniveau von 1%.

Für einfache, mittelschwierige und schwierige Items konnte bei linearer Auswertung in keiner Bedingung eine signifikante Veränderung der Validität nach der Korrektur auf der Basis des Realismusindex gezeigt werden. Eine signifikante Verbesserung der Korrelationskoeffizienten zeigte sich dagegen bei der logarithmischen Auswertung nach der Realismuskorrektur für mittelschwierige und schwierige Items sowohl in der Bedingung mit Items mit zwei Antwortoptionen als auch in der Bedingung mit Items mit drei Antwortoptionen.

Zusammenfassend wurde für die Bedingungen mit Multipler Evaluation und der linearen Auswertung keine Verbesserung der Validität aufgrund der Realismuskorrektur beobachtet. Bei einer logarithmischen Auswertung führte die Korrektur hingegen zu einer zumTeil signifikanten Verbesserung der Validität.

11.3.10 Abbruchquoten

Die im Folgenden als Abbruchquoten bezeichneten Prozentwerte bedeuten, dass von 100% der Teilnehmer, die den jeweiligen Teil des Experiments aufgerufen hatten, der dargestellte Prozentsatz von Teilnehmern nicht zum nächsten Teil des Experiments weiterging, sondern das Experiment an dieser Stelle abbrach. Während des ersten Teils der Testapplikation, dem Fragebogen zur Erhebung der demographischen Daten, beendeten im Durchschnitt über alle Versuchsbedingungen 8,28% (*SD*=0,39%) der Teilnehmer das Experiment. Während der Einführung in das Antwortverfahren Multiple Evaluation brachen im Durchschnitt über die vier Bedingungen 2,22% (*SD*=0,80%) der Teilnehmer das Experiment ab. Während der Einführung in das Multiple-Choice-Verfahren betrug die Abbruchquote 1,10% (*SD*=0,13%).

Wenn die Teilnehmer in den Bedingungen mit Multipler Evaluation die Übungen durchführten, so zeigte sich ein Unterschied in der Abbruchquote in Abhängigkeit von der Anzahl der Antwortoptionen. Hatten die Items zwei Antwortoptionen, so beendeten durchschnittlich 7,35% der Teilnehmer das Experiment. Betrug die Anzahl der Antwortoptionen drei, so war die Abbruchquote mit 15,06% um 104,85% höher (Δ=7,71%). Tabelle 11-33 zeigt die Abbruchquoten während der Übungen für alle Bedingungen mit Multipler Evaluation. Tabelle 11-34 zeigt die Mittelwerte und Standardabweichungen der Abbruchquoten während der Übungen. Dabei wurden die Bedingungen zusammengefasst für die unabhängige

Variable „Anzahl der Antwortoptionen“. Die Teilnehmer der Bedingung mit Multiple-Choice bearbeiteten auch im dritten Experiment keine Übungen.

Tabelle 11-33

Abbruchquoten während der Übungen für alle Bedingungen mit Multipler Evaluation.

Bedingung	% Abbruch während der Übungen
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	6,56
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	13,95
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	8,14
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	16,16

Tabelle 11-34

Mittelwerte und Standardabweichungen der Abbruchquoten während der Übungen. Die Bedingungen wurden für die unabhängige Variable „Anzahl der Antwortoptionen“ zusammengefasst.

Bedingung	% Abbruch während der Übungen-Mittelwert	Standardabweichung
ME-2 Antwortoptionen	7,35	1,12
ME-3 Antwortoptionen	15,06	1,56

Im Folgenden werden die Abbruchquoten während der Bearbeitung der Items des Englischtests betrachtet. Tabelle 11-35 zeigt die Abbruchquoten in Prozent für alle Versuchsbedingungen. Tabelle 11-36 zeigt die Abbruchquoten für die Bedingungen, zusammengefasst für die unabhängigen Variablen „Auswertefunktion“ und „Anzahl der Antwortoptionen“. Wurden alle Bedingungen

Tabelle 11-35

Anteile der Teilnehmer in Prozent, die das Experiment während der Bearbeitung der Items des Englischtests beendeten. Die Tabelle zeigt die Abbruchquoten für alle Versuchsbedingungen.

Bedingungen	% Abbruchquote während der Items
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	10,63
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	13,12
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	21,46
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	26,75
MC-2 Antwortoptionen	12,14
MC-3 Antwortoptionen	15,93

des Experiments mit Items mit zwei Antwortoptionen gemeinsam betrachtet, so zeigte sich eine Abbruchquote während der Bearbeitung der Items von 14,74%. In den Bedingungen, in denen die Items über drei Antwortoptionen verfügten, betrug die Abbruchquote 18,60% und war damit um 26,15% höher (Δ=3,86%). Die Anzahl der Antwortoptionen zeigte also einen Einfluss auf die Anzahl der Teilnehmer, die den Englischtest vorzeitig beendeten.

Auch die Auswertefunktion zeigte einen Einfluss auf die Abbruchquote. Wurde die lineare Auswertefunktion ohne Strafzahlungen verwendet, so beendeten 11,87% der Teilnehmer ihre Teilnahme. Bei der Auswertung durch die logarithmische Funktion war die Abbruchquote mit 23,22% um 95,59% höher (Δ=11,35%).

Tabelle 11-36

Anteile der Teilnehmer in Prozent, die das Experiment während der Bearbeitung der Items des Englischtests beendeten. Die Tabelle zeigt die Abbruchquoten der Bedingungen, zusammengefasst über die unabhängigen Variablen „Auswertefunktion“ und „Anzahl der Antwortoptionen“.

Bedingung	% Abbruchquote während der Items-Mittelwert	Standardabweichung
ME-2 Antwortoptionen	16,05	7,66
ME-3 Antwortoptionen	19,93	9,64
ME-linear	11,87	1,76
ME-logarithmisch	23,22	3,74
2 Antwortoptionen	14,74	10,33
3 Antwortoptionen	18,60	7,20

11.4 Diskussion

Im dritten Experiment wurde der Einfluss der Anzahl der Antwortoptionen auf die Reliabilität und die Validität eines Englischtests anhand von Items mit drei und Items mit zwei Antwortoptionen untersucht. Dabei wurde wiederum das Antwortverfahren Multiple Evaluation im Vergleich zu einem herkömmlichen Multiple-Choice-Verfahren betrachtet. Bei dem Test mit Multipler Evaluation erfolgte die Auswertung anhand einer linearen Funktion ohne Strafzahlungen sowie einer logarithmischen Funktion (Dirkzwager, 2003) mit dem Toleranzfaktor T=3. Damit wurden bei der logarithmischen Auswertung maximal -300 Strafpunkte ausgezahlt.

Durch die Befunde des Experiments konnte für das Antwortverfahren Multiple-Choice die Hypothese bestätigt werden, dass sich die Reliabilität des Englischtests signifikant verringert, wenn die Anzahl der Antwortoptionen je Item von drei auf zwei reduziert wird. Dieses Ergebnis war zu erwarten, denn durch die Reduzierung der Anzahl der Antwortoptionen von drei auf zwei steigt die Wahrscheinlichkeit, ein Item nur durch blindes Raten richtig zu lösen von gerundet 33% auf 50%. Als Folge daraus wurde eine Erhöhung der Varianz der beobachtbaren Testwerte und damit eine Verschlechterung der Reliabilität erwartet. Der Multiple-Choice-Test mit Items mit zwei Antwortoptionen müsste um den Faktor 1,36 verlängert werden, um die Reliabilität des Tests mit Items mit drei Antwortoptionen zu erreichen. Die Berechnung des Faktors der Testverlängerung erfolgte nach der Methode von Spearman-Braun (Bühner, 2006). Wurden für die Bedingungen mit Multiple-Choice die jeweils zwölf Items der drei Schwierigkeitsstufen getrennt ausgewertet, so zeigte sich die deutlichste signifikante Verringerung der Reliabilität, wenn statt drei nur zwei Antwortoptionen dargeboten wurden, bei den mittelschwierigen Items. Diese Verringerung entsprach einer Verlängerung des Tests mit Items mit zwei Antwortoptionen um den Faktor 2,02. Auch bei den einfachen Items war die Reliabilität des Tests mit Items mit zwei Antwortoptionen signifikant geringer, einer Verlängerung des Tests um den Faktor 1,35 entsprechend. Bei schwierigen Items zeigte sich keine signifikante Veränderung der Reliabilität.

In den Bedingungen mit Multipler Evaluation hatten die Teilnehmer die Möglichkeit, ihr Wissen in Form ihrer Antwortsicherheit differenziert wiederzugeben und so auch Nicht- und Teilwissen einzuräumen. Bei der Auswertung durch eine logarithmische Funktion sollten die Teilnehmer zudem

durch die hohen Strafzahlungen bei nur geringen Antwortsicherheiten in eine richtige Antwortoption inzentiviert werden, ihre tatsächlichen subjektiven Antwortsicherheiten unverfälscht zu berichten. Daher sollte die Varianz der beobachtbaren Testwerte des Tests mit Multipler Evaluation nicht im gleichen Maße wie bei Multiple-Choice durch die Anzahl der Antwortoptionen der Items beeinflusst werden. In den Bedingungen mit Multipler Evaluation wurde also erwartet, dass sich die Reliabilität nicht signifikant verschlechtert, wenn Items mit zwei statt drei Antwortoptionen gemessen werden. Diese Hypothese bestätigte sich jedoch nicht. In beiden Auswertebedingungen, der linearen und der logarithmischen, sank die Reliabilität des Tests signifikant, wenn die Anzahl der Antwortoptionen je Item von drei auf zwei reduziert wurde. Der Faktor, um den der Test mit zwei Antwortoptionen verlängert werden müsste, um die Reliabilität des Tests mit drei Antwortoptionen zu erreichen, betrug 1,32 bei einer linearen Auswertung und 1,29 bei einer logarithmischen. Auch die Reliabilität, jeweils ermittelt für einfache, mittelschwierige oder schwierige Items, verringerte sich auf allen Schwierigkeitsstufen signifikant, wenn die Auszahlung der Punkte durch die lineare Funktion ohne Strafzahlungen erfolgte. Wurden die Punkte anhand der logarithmischen Funktion ausgezahlt, so war die Verschlechterung der Reliabilität nur bei mittelschwierigen und schwierigen Items signifikant, jedoch nicht bei den einfachen. Tabelle 11-37 zeigt die Faktoren, um die der Test mit Items mit zwei Antwortoptionen verlängert werden müsste, um die Reliabilität des Tests mit Items mit drei Antwortoptionen zu erreichen. Bei der linearen Auswertung wird eine verfälschte Reproduktion der tatsächlichen subjektiven Antwortsicherheiten nicht

Tabelle 11-37

Faktoren, um die der Test mit zwei Antwortoptionen verlängert werden müsste, um die Reliabilität des Tests mit drei Antwortoptionen zu erreichen.

	Itemschwierigkeit			
Bedingung	alle	einfach (A)	mittel (B)	schwierig (C)
ME-linear 0 bis 100 Punkte	1,32	1,28	2,03	1,28
ME-logarithmisch 300 bis 100 Punkte	1,29	n.s.	1,52	1,24
MC-0 oder 1 Punkt	1,36	1,35	2,20	n.s.

n.s.=der Unterschied in der Reliabilität zwischen den Items mit drei Antwortoptionen und den Items mit zwei Antwortoptionen war nicht signifikant.

bestraft, deshalb ist es für den Teilnehmer eine empfehlenswerte Strategie, immer der plausibelsten Antwortoption seine absolute Antwortsicherheit zuzuordnen, wenn er seine erwartete Punktsumme maximieren will. Die hohen Strafzahlungen bei einer logarithmischen Auswertung scheinen Testteilnehmer jedoch nicht in ausreichendem Maße davon abzuhalten, ihre Antwortsicherheiten verfälscht zu berichten. So hatte auch mit einer logarithmischen Auswertung die Reduzierung der Anzahl der Antwortoptionen eine signifikante Verschlechterung der internen Konsistenz zur Folge. Die geringere Verschlechterung der Reliabilität bei einfachen und schwierigen Items liefert jedoch einen Hinweis darauf, dass die Varianz der beobachtbaren Testwerte bei Multipler Evaluation und einer Auswertung durch eine logarithmische Funktion zumindest ansatzweise deutlicher reduziert werden konnte als mit der linearen Auswertung.

Da die logarithmische Auswertung Teilnehmer nicht ausreichend inzentivierte, ihre tatsächlichen subjektiven Antwortsicherheiten unverfälscht zu berichten, zeigte der Test mit Multipler Evaluation mit einer linearen Auswertung eine höhere Reliabilität als der Test mit einer logarithmischen Auswertung. Es kann angenommen werden, dass die hohen Strafzahlungen, die die Teilnehmer erhielten, wenn sie ihre Antwortsicherheiten verfälscht berichteten, zu einer Erhöhung der Varianz der beobachtbaren Testwerte führten und die Reliabilität des Tests verschlechterten. Im dritten Experiment wurden damit auch die Ergebnisse des zweiten Experiments reproduziert. Auch in diesem Experiment wurde bei einer linearen Auswertung eine höhere Reliabilität des Tests mit Multipler Evaluation beobachtet als bei einer logarithmischen, die hohe Strafzahlungen von maximal -300 Punkten vorsieht.

Die Hypothese, dass die Validität des Multiple-Choice-Tests sinkt, wenn die Items nur zwei statt drei Antwortoptionen haben, konnte durch die Ergebnisse des Experiments nicht bestätigt werden. Wurde die Validität des gesamten Tests betrachtet, so zeigte sich keine signifikante Verringerung in Abhängigkeit von der Anzahl der Antwortoptionen. Die Validität, ermittelt für einfache und mittelschwierige Items, war ebenfalls nicht signifikant verändert. Eine signifikante Verschlechterung der Validität zeigte sich für die Bedingungen mit Multiple-Choice nur für mittelschwierige Items.

In den Bedingungen mit Multipler Evaluation verringerte sich die Validität des gesamten Tests in beiden Auswertebedingungen ebenfalls nicht signifikant, wenn die Anzahl der Items von drei auf zwei reduziert wurde. Wurden die Punkte durch die lineare Funktion ausgezahlt, so zeigte sich eine signifikante Verschlechterung der Validität, wenn die Items der Schwierigkeitsstufen getrennt betrachtet wurden, nur für mittelschwierige Items. In den Bedingungen mit Multipler Evaluation und der Auswertung durch die logarithmische Funktion zeigte die getrennte Betrachtung der Items der drei Schwierigkeitsstufen keine signifikante Veränderung der Validität für einfache und mittelschwierige Items. Bei schwierigen Items war die Validität des Tests mit Items mit zwei Antwortoptionen signifikant höher als die des Tests mit drei Antwortoptionen.

Zu den hier dargelegten Befunden ist anzumerken, dass auch die Validität der Kontrollbedingung mit Multiple-Choice aufgrund der Verringerung der Anzahl der Antwortoptionen der Items nicht signifikant verringert war. Damit stellte die selbsteingeschätzte Fähigkeit der Teilnehmer kein geeignetes Kriterium dar, um die Hypothese zu prüfen, dass die Validität eines Tests mit dem Antwortverfahren Multiple Evaluation nicht durch die Anzahl der Antwortoptionen beeinflusst wird.

Der Vergleich des Tests mit Multipler Evaluation und einer linearen Auswertung zu einer logarithmischen bestätigte auch für die Validität die Ergebnisse des zweiten Experiments. Die Validität des Tests war signifikant höher, wenn die Auswertung mithilfe der linearen Funktion erfolgte, im Vergleich zu einer logarithmischen Auswertung, die hohe Strafzahlungen von bis zu -300 Punkten auszahlte.

Die Betrachtung des individuellen Realismusindex gibt weiteren Aufschluss über das Antwortverhalten der Teilnehmer. Auch im dritten Experiment war nur ein geringer Teil von ihnen annähernd perfekt kalibriert. Die meisten Teilnehmer zeigten *Overconfidence* und berichteten häufig nur geringe Antwortsicherheiten in die richtigen Antworten und somit hohe in die Distraktoren. Es ist davon auszugehen, dass die Teilnehmer in den meisten Fällen nicht fehlinformiert waren, sondern versuchten, trotz Nicht- oder Teilwissen die richtige Antwort zu erraten. Diese Vermutung wird dadurch gestützt, dass die Testteilnehmer bei den Items mit drei Antwortoptionen einen zwar geringen, aber signifikant höheren Realismusindex zeigten als bei den Items mit zwei Antwortoptionen. Es ist zu vermuten, dass sie

durch eine Wahrscheinlichkeit von 50%, die richtige Antwortoption auch bei Nichtwissen zu erraten, eher dazu neigten, ihr Rateglück zu versuchen, als wenn die Wahrscheinlichkeit nur 33% betrug. Die Auswertung durch die logarithmische Funktion mit einer Strafzahlung von maximal -300 Punkten hatte eine zwar signifikante, aber nur geringfügige Verbesserung des durchschnittlichen Realismus-index zur Folge, im Vergleich zu den Bedingungen, bei denen eine lineare Auswertung erfolgte. Daraus kann abgeleitet werden, dass die meisten Teilnehmer ihre tatsächlichen subjektiven Antwortsicherheiten nicht deutlich unverfälschter berichteten und daher keine Reduzierung der Varianz der beobachtbaren Testwerte beobachtet wurde. Daher hatte die Anzahl der Antwortoptionen einen deutlichen Einfluss zumindest auf die Reliabilität des Tests mit Multipler Evaluation.

Um die Varianz der beobachtbaren Testwerte nachträglich zu reduzieren, wurden die Antwortsicherheiten wiederum auf der Basis des individuellen Realismusindex jedes Teilnehmers korrigiert. Die erneute Auszahlung der Punkte für die korrigierten Antwortsicherheiten verbesserte die Reliabilität des gesamten Tests signifikant, sowohl bei einer linearen als auch bei einer logarithmischen Auswertung. Wurden die Punkte durch die lineare Funktion ausgezahlt, so war die Reliabilität nach der Realismuskorrektur, auch wenn jeweils nur die einfachen, mittelschwierigen und schwierigen Items betrachtet wurden, signifikant verbessert, und zwar sowohl für die Items mit zwei als auch für die mit drei Antwortoptionen. Bei der Auswertung durch die logarithmische Funktion war die Reliabilität der einfachen und mittelschwierigen Items ebenfalls signifikant verbessert. Ein inkonsistentes Muster zeigte sich hingegen, wenn die Reliabilität der schwierigen Items betrachtet wurde. Hier war der α-Koeffizient der Items mit zwei Antwortoptionen nach der Korrektur nicht signifikant höher. Der α-Koeffizient der schwierigen Items mit drei Antwortoptionen war nach der Korrektur sogar signifikant verschlechtert.

Die Validität des Englischtests mit der linearen Auswertung konnte durch die Korrektur auf der Basis des individuellen Realismusindex nicht signifikant verbessert werden. Dieser Befund zeigte sich sowohl, wenn alle Items des Tests betrachtet wurden, als auch bei einer getrennten Auswertung der einfachen, mittelschwierigen und schwierigen Items. Für die Bedingungen mit der logarithmischen Auswertung konnte hingegen sowohl für den Test mit Items mit

zwei Antwortoptionen als auch für den Test mit Items mit drei Antwortoptionen eine signifikante Verbesserung der Validität aufgrund der Korrektur beobachtet werden. Wurde die Validität der jeweils zwölf einfachen, mittelschwierigen und schwierigen Items getrennt betrachtet, so fand sich eine signifikante Verbesserung der Validität durch die Korrektur für mittelschwierige und schwierige Items.

Der Vergleich des Tests mit Multipler Evaluation mit dem mit Multiple-Choice zeigte auch im dritten Experiment eine deutliche signifikante Verbesserung der Reliabilität bei der linearen Auswertung. Diese Verbesserung entsprach für den gesamten Test einer Verlängerung des Multiple-Choice-Tests um den Faktor 1,47 für die Items mit drei Antwortoptionen und um den Faktor 1,36 für die Items mit zwei Antwortoptionen. Die Verbesserung der Reliabilität war mit der Auswertung durch die lineare Funktion sowohl für die Items mit zwei als auch für die mit drei Antwortoptionen für einfache, mittelschwierige und schwierige Items signifikant. Die höchste Verbesserung der Reliabilität zeigte sich dabei wiederum für schwierige Items. Sie entsprach einer Verlängerung des Multiple-Choice-Tests um den Faktor 1,58 für die Testbedingung mit Items mit zwei Antwortoptionen und für die mit drei Antwortoptionen um den Faktor 1,80. Wurden die Punkte durch die logarithmische Funktion ausgezahlt, so zeigte sich hingegen keine signifikante Verbesserung der Reliabilität des gesamten Tests gegenüber dem Test mit Multiple-Choice. Nur für schwierige Items mit drei Antwortoptionen war die Reliabilität signifikant höher. Diese Verbesserung entsprach einer Verlängerung des Multiple-Choice-Tests um den Faktor 1,35. Tabelle 11-38 zeigt die Faktoren, um die der Englischtest mit Multiple-Choice verlängert werden müsste, um die Reliabilität der Bedingungen mit Multipler

Tabelle 11-38

Faktoren, um die der Englischtest mit Multiple-Choice verlängert werden müsste, um die in den Bedingungen mit Multipler Evaluation erzielte Reliabilität ebenfalls zu erreichen.

	Schwierigkeit			
Bedingung	alle	einfach (A)	mittel (B)	schwierig (C)
ME-linear 0 bis 100 Punkte-2 Antwortoptionen	1,36	1,42	1,44	1,58
ME-linear 0 bis 100 Punkte-3 Antwortoptionen	1,47	1,33	1,42	1,80
ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen	n.s.	n.s.	n.s.	1,18
ME-logarithmisch -300 bis 100 Punkte-3 Antwortoptionen	n.s.	n.s.	n.s.	1,35

n.s.=der Unterschied in der Reliabilität war nicht signifikant.

Evaluation zu erreichen, für den gesamten Test und für die einfachen, mittelschwierigen und schwierigen Items.

Die Hypothese, der Test mit Multipler Evaluation würde gegenüber dem Test mit Multiple-Choice eine signifikant höhere Validität zeigen, bestätigte sich für den gesamten Englischtest bei einer linearen Auswertung nicht. Auch die getrennte Betrachtung der einfachen, mittelschwierigen und schwierigen Items zeigte keine signifikante Verbesserung der Validität der Multiplen Evaluation mit der linearen Auswertung im Vergleich zu Multiple-Choice. Bei der Auswertung durch eine logarithmische Funktion war die Validität des Tests sowohl mit zwei, als auch mit drei Antwortoptionen signifikant gegenüber der entsprechenden Bedingung mit Multiple-Choice verschlechtert. Die Verschlechterung der Validität zeigte sich hierbei besonders für mittelschwierige und schwierige Items.

Die Abbruchquoten des dritten Experiments waren geringfügig höher, wenn die Items drei Antwortoptionen hatten, gegenüber denen mit zwei Optionen. Die etwas höhere Schwierigkeit der Items mit drei Antwortoptionen könnte die Ursache dafür sein. Auch im dritten Experiment zeigte sich ein deutlicher Einfluss der Auswertefunktion auf die Abbruchquoten des Experiments. In den Bedingungen mit der logarithmischen Funktion und damit einer Strafzahlung von bis zu -300 Punkten brachen deutlich mehr Teilnehmer das Experiment ab als in den Bedingungen, in denen die Punkte durch die lineare Funktion ohne Strafpunkte ausgezahlt wurden. Daraus lässt sich ableiten, dass die hohen Strafzahlungen bei geringen Antwortsicherheiten Teilnehmer frustrierten und sie den Test deshalb vorzeitig beendeten.

12 Abschließende Diskussion

Das Ziel dieser Arbeit war die Untersuchung des bisher noch wenig erforschten Antwortbewertungsverfahrens (Multiple Evaluation) im Vergleich zu einem herkömmlichen Antwortwahlverfahren (Multiple-Choice). Im Folgenden werden die Ergebnisse aller drei Experimente zusammengefasst und deren Bedeutung für den Einsatz von Multipler Evaluation in der Wissensdiagnostik diskutiert.

Der Unterschied zwischen einem Test mit Multipler Evaluation und einem Test mit Multiple-Choice liegt darin, dass ein Teilnehmer sich nicht nur für eine am wahrscheinlichsten richtige Antwortoption entscheiden muss, sondern seine Antwortsicherheit in jede der dargebotenen Antwortoptionen angibt. Das hat den Vorteil, dass der Teilnehmer auch Nicht- und Teilwissen einräumen kann. Bisher war noch nicht empirisch untersucht, wie differenziert Antwortsicherheit dabei erhoben werden muss, damit diese Vorteile der Multiplen Evaluation im Vergleich zu Multiple-Choice messbar werden. Im ersten Experiment wurde der Einfluss einer prozentgenauen Erfassung von Antwortsicherheit mithilfe von Schiebereglern im Vergleich zu einem Antwortdreieck mit nur 16 Antwortmöglichkeiten auf die Reliabilität eines Englischtests geprüft. Die prozentgenaue Erfassung führte dabei nicht zu einer Verbesserung der Reliabilität. Als Ursache dieses Befundes ist anzunehmen, dass die Genauigkeit, mit der Teilnehmer in der Lage sind, ihre Antwortsicherheit zu differenzieren, begrenzt ist, so dass eine diskrete Anzahl von Antwortmöglichkeiten eine ausreichende Genauigkeit bieten (Leclerq, 1983; Paul, 1993). Das Einstellen mehrerer Schieberegler ist zudem aufwendig und daher zeitintensiv, während beim Dreieck durch die Auswahl eines Feldes die simultane Angabe von Antwortsicherheit in drei Antwortoptionen erfolgt. Es wurde daher beobachtet, dass die Durchführungszeit des Tests signifikant mit den Schiebereglern im Vergleich zum Dreieck verlängert war. Die Untersuchung der beiden Antwortinstrumente lässt daher den Schluss zu, dass das ökonomischere Antwortdreieck eine ausreichend genaue Erhebung von Antwortsicherheit ermöglicht und daher ein geeignetes Antwortinstrument zu Erhebung von Antwortsicherheit ist.

Kritisch muss gesehen werden, dass sich die schon von Ebel (1968) kritisierte verlängerte Bearbeitungszeit eines Tests mit Multipler Evaluation im Vergleich zu

Multiple-Choice, auch unter dem Einsatz eines Computers bestätigte. Dabei war die Durchführungszeit sowohl für die Schieberegler als auch für das Antwortdreieck verlängert.

Als eine weitere Fragestellung wurde untersucht, ob eine Verbesserung der Reliabilität eines Tests mit Multipler Evaluation und einer logarithmischen Auswertung dadurch erreicht werden kann, dass die Teilnehmer über die zu erwartende (*Feedforward*) und die tatsächliche Auszahlung (*Feedback*) informiert werden. Die Ergebnisse geben keinen Hinweis darauf, dass die Reliabilität eines Tests mit Multipler Evaluation durch eine Information der Teilnehmer über die Auszahlung verbessert werden kann. Dies lässt die Schlussfolgerung zu, dass ein Test mit Multipler Evaluation mit einem Antwortdreieck ohne Information des Teilnehmers über die Auszahlung in der Wissensdiagnostik eingesetzt werden kann. Mit dem Dreieck als Antwortinstrument wäre daher auch ein Test mit Multipler Evaluation in Form einer Papier-und-Bleistift-Version denkbar. Die Schieberegler hingegen können nicht als Papier-und-Bleistift-Version realisiert werden, da sonst nicht sichergestellt werden kann, dass ein Teilnehmer immer in der Summe 100% Antwortsicherheit in die Antwortoptionen angibt. Bei Trainingsexperimenten, bei denen Teilnehmer es lernen sollen, ihre Antwortsicherheiten differenziert anzugeben, sind die Schieberegler dem Dreieck vorzuziehen, da diese die Genauigkeit, mit der die Testteilnehmer ihre Antwortsicherheit berichten können, nicht auf 16 Möglichkeiten beschränken. Außerdem wurde im ersten Experiment dieser Arbeit gezeigt, dass der durchschnittliche Realismusindex der Teilnehmer bei Verwendung der Schieberegler signifikant höher war als mit dem Antwortdreieck. Ein weiterer Vorteil der Schieberegler gegenüber dem Dreieck ist, dass die Anzahl der Antwortoptionen variiert werden kann, während das Dreieck auf genau drei Optionen beschränkt ist. Die Befunde des dritten Experiments geben jedoch Hinweise darauf, dass, wie bei einem Test mit Multiple-Choice, auch bei einem Test mit Multipler Evaluation Items mit drei Antwortoptionen zumindest eine höhere Reliabilität liefern als Items mit zwei Antwortoptionen.

Eine logarithmische Auswertung stellt bei einem Test mit Multipler Evaluation sicher, dass ein Teilnehmer sein Ergebnis nur dann maximieren kann, wenn er seine tatsächlichen subjektiven Antwortsicherheiten völlig unverfälscht berichtet. Je höher der Toleranzfaktor in der Auswertung ist, desto höher sind die Strafzahlungen, die

ein Teilnehmer erhält, wenn er nur eine geringe Antwortsicherheit in die richtige Antwortoption angibt. Der Toleranzfaktor gibt dabei an, wie viele Items ein Teilnehmer vollkommen richtig beantworten muss, um den Punkteverlust durch ein mit 0% Antwortsicherheit völlig falsch beantwortetes Item wieder aufzuwiegen. In allen drei Experimenten erfolgte die Auswertung anhand einer logarithmischen Funktion nach Dirkzwager (2003). Im ersten Experiment wurde dabei der Toleranzfaktor T=1 gewählt. Im zweiten Experiment wurde die Auswertung mit den Toleranzfaktoren T=0,5, T=1 und T=3 untersucht. Bei der logarithmischen Auswertung im dritten Experiment betrug der Toleranzfaktor T=3. Im zweiten und dritten Experiment wurde zudem die Auswertung anhand einer linearen Funktion, die die prozentuale Antwortsicherheit in die richtige Antwort als Punkte auszahlt, untersucht. Bei einer linearen Auswertung wird ein Teilnehmer nicht durch hohe Strafzahlungen inzentiviert, seine tatsächlichen subjektiven Antwortsicherheiten unverfälscht zu berichten, sondern es ist für ihn sogar die sinnvollste Teststrategie, immer der am wahrscheinlichsten richtigen Antwortoption seine absolute Antwortsicherheit zuzuordnen, wenn er sein Testergebnis maximieren will. Die Frage, welche Auswertefunktion ein Testleiter für Multiple Evaluation verwenden sollte, kann jedoch anhand der Befunde der drei Experimente nicht eindeutig beantwortet werden. Die psychometrische Güte eines Tests mit Multipler Evaluation und einer Auswertung durch eine logarithmische Funktion nach Dirkzwager (2003) wird beeinflusst durch die Höhe des Toleranzfaktors in der Auswertung und die Schwierigkeit der Items. Je höher der Toleranzfaktor, d.h. je höher also die Strafzahlung bei einer nur geringen Antwortsicherheit in die richtige Antwortoption war, desto mehr verschlechterten sich die Reliabilität und die Validität des Englischtests. Damit stehen die Befunde im Widerspruch zu den beispielsweise von Dirkzwager (2003), Shuford, Albert und Massengill (1966) sowie Shuford und Brown (1975) abgeleiteten Vorhersagen, dass die Auswertung anhand einer logarithmischen Funktion Teilnehmer zu einer unverfälschten Reproduktion ihrer tatsächlichen subjektiven Antwortsicherheiten inzentiviere und die Güte eines Tests daher durch eine solche Auswertung verbessert werden könne. Eine wesentliche Ursache dieser Befunde scheint in der nicht perfekten Kalibrierung vieler Teilnehmer zu liegen, denn ein hoher Anteil von ihnen zeigte *Overconfidence*. Offenbar berichteten also viele Teilnehmer trotz der zu erwartenden hohen Strafzahlungen Antwortsicher-

heiten, die nicht ihren tatsächlichen subjektiven entsprachen. So konnte die Varianz der beobachtbaren Testwerte bei der Auswertung durch die logarithmische Funktion nach Dirkzwager (2003) nicht ausreichend reduziert werden, um die Reliabilität und die Validität deutlich zu verbessern. Stattdessen erhielten die Testteilnehmer teilweise hohe Strafzahlungen, die ihrerseits die psychometrische Qualität des Tests negativ beeinflussten. Die höheren Abbruchquoten während des Englischtests, in Abhängigkeit von der Höhe der Strafzahlungen in der Auswertung, lassen zudem auf einen motivationalen Einfluss der Strafzahlungen auf die Teilnehmer schließen.

Die Befunde des zweiten und dritten Experiments zeigten dagegen, dass die höchste psychometrische Güte des Englischtests erreicht wurde, wenn die Auswertung anhand einer linearen Funktion ohne Strafzahlungen erfolgte. Unter dieser Auswertung ist es jedoch für einen Teilnehmer die beste Strategie, immer der plausibelsten Antwortoption eine absolute Antwortsicherheit zuzuordnen, wenn er seine Punkte maximieren will. Es ist jedoch zu vermuten, dass die meisten Teilnehmer in den Experimenten dieser Arbeit, auch mit der Auswertung durch die lineare Funktion, noch keine Teststrategien entwickelt hatten, da sie das Antwortverfahren zum ersten Mal verwendeten. Würde Multiple Evaluation jedoch zu einem häufig eingesetzten Antwortverfahren, so wäre davon auszugehen, dass Teilnehmer über Wissen um die beste Teststrategie verfügen. Wenn Teilnehmer immer eine absolute Antwortsicherheit in die plausibelste Antwortoption berichten würden, dann würde Multiple Evaluation mit einer linearen Auswertung zu Multiple-Choice reduziert.

Das Dilemma, dass die lineare Funktion eine verfälschte Reproduktion der tatsächlichen subjektiven Antwortsicherheiten belohnt, während die logarithmische Auswertung mit hohen Strafzahlungen keine optimale Verbesserung der Testgütekriterien gegenüber Multiple-Choice bewirkt, könnte jedoch durch eine nachträgliche Korrektur der berichteten Antwortsicherheiten auf der Basis des individuellen Realismusindex eines Testteilnehmers gelöst werden. In dieser Arbeit konnte gezeigt werden, dass zumindest die Reliabilität eines Tests durch die Korrektur auf der Basis des Realismusindex deutlich signifikant verbessert werden kann. Es wäre außerdem möglich, den Teilnehmern eine Rückmeldung darüber zu geben, wie viele Punkte sie durch eine verfälschte Reproduktion ihrer tatsächlichen subjektiven Antwortsicherheiten verloren haben. Jans und Leclercq (1997) sowie

Shuford und Brown (1975) argumentierten, dass eine solche Rückmeldung den Lernprozess unterstützen könne. Teilnehmer können so möglicherweise lernen, ihre Kalibrierung dauerhaft zu verbessern.

Anhand der Experimente dieser Arbeit wurde auch gezeigt, dass Multiple Evaluation in Form einer Webanwendung, bei der die Teilnehmer den Test ohne Hilfestellung durch einen Testleiter durchführen, einsetzbar ist.

13 Ausblick

In dieser Arbeit wurde gezeigt, dass durch den Einsatz des Antwortbewertungsverfahrens Multiple Evaluation im Vergleich zu dem herkömmlichen Multiple-Choice-Verfahren in Abhängigkeit von der Schwierigkeit der Items und der verwendeten Auswertefunktion die Reliabilität und zum Teil auch die Validität eines Tests verbessert werden können. Jedoch sind noch viele Fragen zur Multiplen Evaluation offen, die im Rahmen dieser Arbeit nicht untersucht werden konnten. Als Testmaterial wurde in dieser Arbeit ausschließlich ein homogener Englischtest verwendet. In weiteren Experimenten sollte daher geprüft werden, ob die Ergebnisse auch auf andere, evtl. heterogene Testmaterialien generalisiert werden können. Da die Ergebnisse bezüglich der Validität in dieser Arbeit inkonsistent waren, sollten weitere Validitätskriterien, wie beispielsweise ein Paralleltest, herangezogen werden. Wichtig wäre es auch, das Verfahren in einem realen Testumfeld zu prüfen, in dem das Ergebnis Konsequenzen für den Teilnehmer hat. Die Durchführung mit dem Dreieck als Papier-und-Bleistift-Version könnte zeigen, ob das Antwortverfahren Multiple Evaluation auch ohne Computer im Testalltag eingesetzt werden kann. Die Untersuchung weiterer admissibler Auswertefunktionen, wie z.B. der sphärischen, die keine Strafzahlungen beinhaltet, sollte fortgesetzt werden. Denn beim Einsatz einer logarithmischen Auswertefunktion stellt sich die Frage, ob die Bestrafung durch Punktabzug aus lerntheoretischer Sicht ein geeignetes Vorgehen ist, da immer die Gefahr besteht, dass negative Reaktionen, wie z.B. Testangst, dadurch verstärkt werden (Bodenmann, Perrez, Schär & Trepp, 2004). Ein umfangreiches Forschungsgebiet ist auch die Messung des Realismusindex als ein Maß der Güte der Kalibrierung von Teilnehmern. Eine wichtige Fragestellung wäre dabei, welche Anzahl von Items erforderlich ist, um eine valide Aussage über die Kalibrierung der Teilnehmer treffen zu können und welchen Einfluss die Schwierigkeit der Items dabei zeigt. Eine weitere Forschungsfrage ist, ob die Güte der Kalibrierung der Teilnehmer auf verschiedene, auch heterogene Testmaterialien generalisiert werden kann.

Zusammenfassend zeigen die Ergebnisse dieser Arbeit, dass Multiple Evaluation eine einsatzfähige Alternative zu Multiple-Choice darstellt. Bevor das Antwortverfahren als Standardinstrument der Wissensdiagnostik eingesetzt werden

kann, ist jedoch noch eine intensivere Erforschung des Verfahrens erforderlich.

14 Literatur

Abedi, J. & Bruno, J. E. (1989). Test-retest reliability of computer based MCW-APM test scoring methods. *Journal of Computer-Based Instruction, 16, 29-35.*

Abedi, J. & Bruno, J. E. (1993). Concurrent validity of the information referenced testing format using MCW-APM scoring methods. *Journal of Computer-Based Instruction 20, 21-25.*

Adams, J. K. & Adams, P. A. (1961). Realism of confidence judgements. *Psychological Review, 68, 33-45.*

Akeroyd, F. M. (1982). Progress in multiple choice scoring methods. *Journal of further and higher Education, 6, 87-90.*

Anderson, R. I. (1982). Computer-based confidence testing: Alternatives to conventional, computer-based multiple-choice testing. *Journal of Computer-Based Instruction, 9, 1-9.*

Bandilla, W. (2002). Web surveys – An appropriate mode of data collection for social sciences? In: Batinic, B., Reips, U. & Bosnjak, M. (2002). *Online Social Sciences, (S. 1-6).* Göttingen: Hogrefe & Huber.

Beckmann, J. E. & Beckmann, N. (2005). Effects of feedback on performance and response latencies in untimed reasoning tests. *Psychology Science, 47, 262-278.*

Ben-Simon, A., Budescu, D.V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple choice tests. *Applied Psychological Measurement, 21, 65-88.*

Bickel, J. E. (2007). Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis, 4, 49-65.*

Bodenmann, G., Perrez, M., Schär, M. & Trepp, A. (2004). *Klassische Lerntheorien*. Bern: Huber.

Bokhorst, F. D. (1986). Confidence weighting and the validity of achievement tests. *Psychological Reports, 59, 383-386.*

Bortz, J. (2005). *Statistik.* (6. Auflage). Heidelberg: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* (4., überarbeitete Auflage). Heidelberg: Springer.

Bosnjak, M. (2002). *(Non) Response bei Webbefragungen.* Mannheim: Unveröffentliche Dissertation Universität Mannheim.

Bradbard, D. A. & Green, S. B. (1986). Use of the Coombs elimination procedure in classroom tests. *Journal of Experimental Education, 54, 68-72.*

Bradbard, D. A., Parker, D. F. & Stone, G. L. (2004). An alternative multiple choice scoring procedure in a macroeconomics course. *Decision Science Journal of Innovative Education, 2, 11-26.*

Brown, T. A. & Shuford, E. H. (1973). Quantifying uncertainty into numerical probabilities for the report of intelligence. *A report prepared for defence advanced research project agency, R-1185-ARPA July 1973, ARPA Order No.: 189-1 3G10 Tactical Technology.*

Bruno, J. E. (1986). Assessing the knowledge base of students: An information theoretic approach for testing. *Measurement and Evaluation in Counselling and Development, 19, 116-130.*

Bruno, J. E. & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55, 959-966.*

Buchanan, T. & Smith, J. L. (1999). Using the internet for psychological research: Personality testing on the world wide web. *British Journal of Psychology, 90, 125-144.*

Budescu, D. & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30, 277-291.*

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. (2., aktualisierte Auflage). München: Pearson.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26, 41-50.*

Coombs, J. E. & Womer F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16, 13-37.*

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

Curlette, W. L. (1978). Demonstration of response strategies in a confidence-testing procedure. *Psychological Reports, 43, 479-485.*

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *The British Journal of Mathematical and Statistical Psychology, 18, 87-123.*

Delgado, A. R. & Prieto, G. (2003). The effect of item feedback on multiple-choice test responses. *British Journal of Psychology, Health and Medical Complete, 94, 73-85.*

Diamond, J. J. (1975). A preliminary study of the reliability and validity of a scoring procedure based upon confidence and partial information. *Journal of Educational Measurement, 12, 129-133.*

Dirkzwager, A. (1993). A computer environment to develop valid and realistic predictions. *Item banking/interactive testing and self-assessment;[proceedings of the NATO Advanced Research Workshop of Item Banking: Interactive Testing and Self-Assessment, held in Liège, Belgium, October, 27-31].*

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

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

Dorsch, F., Häcker, H. & Stapf, K. (1992). *Dorsch Psychologisches Wörterbuch.* (Nachdruck der 11. ergänzten Auflage von 1987). Bern: Huber.

Dressel, P. L. & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 36, 301-310.*

Ebel, R. L. (1968). Valid confidence testing-demonstration kit. *Journal of Educational Measurement, 5, 353-354.*

Echternacht, G. J. (1972). The use of confidence testing in objective tests. *Review of Educational Research, 42, 217-236.*

Feldt, L. S., Woodruff, D. L. & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11, 93-103.*

Fisseni, H. J. (1997). *Lehrbuch der psychologischen Diagnostik.* (2., überarbeitete und erweiterte Auflage). Göttingen: Hogrefe.

Fischoff, B., Slovic, P. & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3, 552-564.*

Gosling, S. D., Vazire, S., Srivastava, S. & John, O. P. (2004). Should we trust web-based studies? *American Psychologist, 59, 93-104.*

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12, 109-113.*

Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement, 53, 999-1010.*

Hambleton, R. K., Roberts, D. M. & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement, 7, 75-82.*

Harsch, C. (2005*). Der Gemeinsame europäische Referenzrahmen für Sprachen: Leistung und Grenzen*. Augsburg: Unveröffentlichte Dissertation, Philologisch-Historische Fakultät der Universität Augsburg.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association, 70, 271-289.*

Holmes, P. (2002). *Multiple evaluation vs multiple choice as testing paradigm.* Published master's thesis. Twente: University Twente.

Jans, V. & Leclercq, D. (1997). Metacognitive realism: A cognitive style or a learning strategy? *Educational Psychology, 17, 101-109.*

Kansup, W. & Hakstian, R. A. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement, 12, 219-230.*

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica, 77, 217-273.*

Koele, P., De Boo, R. & Verschure, P. (1987). Scoring rules and probability testing. *Bulletin of Psychonomic Society, 25, 280-282.*

Koehler, R. A. (1971). A comparison of the validities of conventional choice testing and various confidence marking procedures. *Journal of Educational Measurement, 8, 297-303.*

Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.*

Krantz, J. H. & Dalal, R. (2000). Validity of web-based psychological research. In: Birnbaum, M. H. (2000). *Psychological Experiments on the Internet (*pp89-117). San Diego: Academic Press.

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

Kulhavy, R. W., Yekovich, F. R. & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68, 522-528.*

Lautenschlager, G. J. (1989). ALPHATST: Testing the differences in values of coefficient alpha. *Applied Psychological Measurement, 13, 284.*

Leclerq, D. (1983). Confidence Marking. *Evaluation in Education, 6, 191-287.*

Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26, 149-171.*

Locke, S. D. & Gilbert, B. O. (1995). Method of psychological assessment, self-disclosure, and experiential differences: A study of computer, questionnaire, and interview assessment formats. *Journal of Social Behavior and Personality, 10, 255-263.*

May, R. S. (1987). *Realismus von subjektiven Wahrscheinlichkeiten*. Frankfurt: Lang.

Metcalf, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2, 100-110.*

Michael, J. J. (1968). The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement, 5, 307-314.*

Murphy, A. H. & Winkler, R. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica, 34, 273-286.*

Oakes, W. (1972). External validity and the use of real people as subjects. *American Psychologist, 27, 959-962.*

Paul, J. (1993). Alternative assessment for software engineering education. *People and Computer, 9, 463-472.*

Prihoda, T. J., Pinckard, R. N. & McMahan, C. A. & Jones, A. C. (2005). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education, 70, 387-386.*

Reips, U. (2000a). The web experiment method: Advantages, disadvantages, and solutions. In: Birnbaum, M. H. (2000). *Psychological experiments on the internet* (pp89-117). San Diego: Academic Press.

Reips, U. (2000b). *Das psychologische Experimentieren im Internet.* In: Batinic, B. (2000). Internet für Psychologen. (2., überarbeitete und erweiterte Auflage) (S. 319-343). Göttingen: Hogrefe.

Reips, U. (2001). The web experimental psychology lab: Five years of data collection on the internet. *Behavior Research Methods, Instruments & Computers, 33, 201-211.*

Reips, U. (2003). Webexperimente - Eckpfeiler der Online-Forschung. In: Theobald, A., Dreyer, M. & Starsetzki, T. [Hrsg.] *Online Marktforschung* (S. 73-89*). (*2. Auflage). Wiesbaden: Gabler.

Reips, U. (2005). Datenautobahnen nutzen. Formen der Internet gestützten Datenerhebung. *Psychoscope, 8, 5-9.*

Rippey, R. M. (1968a). A 7094 Fortran IV program for scoring and analyzing probabilistic tests. *Behavioral Science: Journal of the Society for the Systems Science, 13, 424.*

Rippey, R. M. (1968b). Probabilistic testing. *Journal of Educational Measurement, 5, 211-215.*

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

Rippey, R. M. (1979). Improving the reliability and validity of confidence-scored tests by adjusting for realism. *Evaluation and the Health Professions, 1, 100-109.*

Rippey, R. M. & Voytovich, A. E. (1982). Adjusting confidence tests for realism. *Evaluation and the Health Professions, 5, 71-85.*

Rippey, R. M. & Voytovich, A. E. (1983). Linking knowledge, realism and diagnostic reasoning by computer-assisted confidence testing. *Journal of Computer-Based Instruction, 9, 88-97.*

Rippey, R. M. & Voytovich, A. E. (1985). Anomalous responses on confidence-scored tests. *Evaluation and the Health Professions, 8, 109-120.*

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

Romberg, T. A., Shepler, J. L. & Wilson, J. W. (1970). Three experiments involving probability measurement procedures with mathematics test items. *Wisconsin University, Madison. Research and Development Center for Cognitive Learning. Sponsors: Office of Education (DHEW), Washington, D.C., Bureau of Research, June 1970, Report No TR 129, ERIC Journal Number ED044315.*

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion.* (2., vollständig überarbeitete und erweiterte Auflage). Bern: Huber.

Rowley, G. L. & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14, 15-22.*

Schmidt, W. C. (1997). World-wide-web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments and Computers, 29, 274-279.*

Sharp, G. L., Cutler, B. L. & Penrod, S. D. (1988). Performance Feedback improves the resolution of confidence judgements. *Organizational Behavior and Human Decision Processes, 42, 271-283.*

Shuford, E. H. & Massengill, H. E. (1966). Decision-theoretic psychometrics: An interim report. *The first semiannual technical report (which covers the period May 1966 through October 1966) of work performed under contract number AF 49(638)-1744, ARPA Order Number 833, by the Shuford-Massengill Corporation, P.O. Box 26, Lexington, Massachusetts, 02173.*

Shuford, E. H., Albert, A. & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika, 31, 125-145.*

Shuford, E. H. (1969). The logic of SCoRule testing. *Based on a paper read at the 11th Annual Conference of the Military Testing Association, 15-19 September 1969, hosted by U.S. Coast Guard Training Center, Governors Island, New York.*

Shuford, E. H. (1971). An evaluation of the SCoRule method for the administration of multiple-choice tests within the atlantic fleet training command. *Lexington: Shuford-Massengill Corporation, P.O. Box 26, Lexington, Massachusetts, 02173.*

Shuford, E. H. & Brown, T. A. (1975). Elicitation of probabilities and their assessment. *Instructional Science, 4, 137-188.*

Shuford, E. H. (1993). Scoring systems, studying and success. *A keynote address at orderwijs research dagen-MECC Maastricht, 26-27-28 mei.*

Smart, R. (1966). Subject selection bias in psychological research. *The Canadian Psychologist, 7, 115-121.*

Toda, M. (1963). Measurement of probability distributions. *Division of Mathematical Psychology, Institute for Research, State College Pennsylvania, Report No. 3, April.*

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *American Statistical Association Journal, 64, 1073-1078.*

Wood, R. (1991). *Assessment and testing.* Cambridge: University Press.

Anhang

A. Daten des ersten Experiments

A.1 Punktauszahlungen

Tabelle A.1-1: Mittelwert und Standardabweichung der Punktauszahlung der Bedingungen mit Multipler Evaluation und dem Antwortdreieck, N=804 je Bedingung.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-Dreieck ohne Auszahlungsinformation Mittelwert	Standardabweichung	ME-Dreieck mit Auszahlungsinformation Mittelwert	Standardabweichung
1 bis 6	A1	3	94,0	33,6	95,3	29,9
1 bis 6	A1	4	90,6	38,6	94,2	30,1
1 bis 6	A1	9	87,5	43,2	86,3	47,3
1 bis 6	A1	13	70,1	66,8	72,8	65,4
1 bis 6	A1	15	76,9	57,2	79,4	55,3
1 bis 6	A1	19	90,2	36,8	92,2	33,6
7 bis 12	A2	35	69,6	63,7	68,8	66,7
7 bis 12	A2	36	56,0	74,0	57,1	76,8
7 bis 12	A2	37	73,5	58,5	72,8	61,8
7 bis 12	A2	38	79,4	51,8	79,5	54,1
7 bis 12	A2	39	65,1	65,3	69,4	64,0
7 bis 12	A2	56	68,7	62,7	75,7	58,0
13 bis 18	B1	67	33,1	83,2	39,1	82,8
13 bis 18	B1	72	30,8	87,6	35,3	87,1
13 bis 18	B1	73	58,0	70,9	64,5	67,7
13 bis 18	B1	75	51,1	70,5	57,0	70,5
13 bis 18	B1	76	42,3	81,3	45,1	80,7
13 bis 18	B1	79	38,1	81,5	42,5	81,2
19 bis 24	B2	82	45,8	75,0	52,3	73,2
19 bis 24	B2	85	33,6	79,4	28,8	82,7
19 bis 24	B2	89	28,7	82,0	30,3	84,4
19 bis 24	B2	90	11,4	80,1	10,2	84,0
19 bis 24	B2	91	-1,6	82,3	2,9	86,4
19 bis 24	B2	95	10,7	82,6	14,8	81,7
25 bis 30	C1	98	26,1	78,0	19,4	79,2
25 bis 30	C1	100	20,6	66,7	17,8	67,5
25 bis 30	C1	101	2,0	84,9	-1,3	85,5
25 bis 30	C1	110	3,3	72,9	-3,9	76,0
25 bis 30	C1	111	9,6	76,2	14,5	76,3
25 bis 30	C1	113	5,0	80,0	7,6	81,2
31 bis 36	C2	125	10,5	58,6	13,1	60,0
31 bis 36	C2	134	-14,1	73,2	-13,5	74,0
31 bis 36	C2	143	-3,9	54,1	-7,5	60,1
31 bis 36	C2	144	-14,4	72,6	-9,7	75,2
31 bis 36	C2	153	1,0	78,2	0,4	75,9
31 bis 36	C2	157	8,4	75,6	12,9	75,7

Tabelle A.1-2: Mittelwert und Standardabweichung der Punktauszahlung der Bedingungen mit Multipler Evaluation und den Schiebereglern.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-Schieberegler ohne Auszahlungsinformation Mittelwert	Standardabweichung	ME-Schieberegler mit Auszahlungsinformation Mittelwert	Standardabweichung
1 bis 6	A1	3	94,9	30,0	96,8	23,6
1 bis 6	A1	4	88,6	41,6	92,6	32,7
1 bis 6	A1	9	85,3	44,5	87,3	42,2
1 bis 6	A1	13	66,9	69,3	70,9	64,0
1 bis 6	A1	15	76,6	53,9	80,4	50,5
1 bis 6	A1	19	88,6	37,7	91,3	32,6
7 bis 12	A2	35	61,9	71,2	69,1	64,5
7 bis 12	A2	36	55,2	73,7	60,1	71,2
7 bis 12	A2	37	75,3	55,3	75,4	55,6
7 bis 12	A2	38	78,7	51,1	83,1	45,3
7 bis 12	A2	39	62,4	65,4	69,1	61,4
7 bis 12	A2	56	72,4	57,3	76,8	52,2
13 bis 18	B1	67	38,0	79,6	37,5	77,9
13 bis 18	B1	72	33,6	84,8	39,7	81,3
13 bis 18	B1	73	57,8	71,4	66,1	64,0
13 bis 18	B1	75	56,5	67,1	59,5	63,5
13 bis 18	B1	76	39,7	82,0	44,7	77,2
13 bis 18	B1	79	44,4	78,6	44,8	77,8
19 bis 24	B2	82	42,4	77,2	50,1	70,5
19 bis 24	B2	85	33,2	80,1	33,3	76,2
19 bis 24	B2	89	23,8	82,9	31,0	80,6
19 bis 24	B2	90	11,1	78,1	16,9	76,9
19 bis 24	B2	91	-0,1	85,3	1,1	82,2
19 bis 24	B2	95	12,0	78,9	15,4	77,3
25 bis 30	C1	98	27,7	72,3	27,8	68,3
25 bis 30	C1	100	19,6	65,7	20,7	61,2
25 bis 30	C1	101	4,1	81,2	14,0	74,8
25 bis 30	C1	110	1,8	70,8	1,2	68,1
25 bis 30	C1	111	11,0	72,9	20,2	68,9
25 bis 30	C1	113	-1,8	78,5	9,9	74,7
31 bis 36	C2	125	12,0	56,7	11,2	54,2
31 bis 36	C2	134	-11,3	67,9	-5,1	64,4
31 bis 36	C2	143	-3,2	52,8	-2,9	48,5
31 bis 36	C2	144	-10,8	66,7	-3,2	62,5
31 bis 36	C2	153	-1,2	71,8	1,9	67,7
31 bis 36	C2	157	7,8	71,2	8,6	64,8

N=804 je Bedingung.

Tabelle A.1-3: Mittelwert und Standardabweichung der Bedingungen mit Multiple-Choice.

Position im Test	GER-Stufe	Item Datenbank Nr.	MC ohne Auszahlungsinformation Mittelwert	MC ohne Auszahlungsinformation Standardabweichung	MC mit Auszahlungsinformation Mittelwert	MC mit Auszahlungsinformation Standardabweichung
1 bis 6	A1	3	0,980	0,140	0,986	0,116
1 bis 6	A1	4	0,932	0,253	0,954	0,210
1 bis 6	A1	9	0,893	0,309	0,922	0,269
1 bis 6	A1	13	0,795	0,404	0,812	0,391
1 bis 6	A1	15	0,874	0,332	0,900	0,300
1 bis 6	A1	19	0,944	0,230	0,955	0,207
7 bis 12	A2	35	0,811	0,392	0,840	0,367
7 bis 12	A2	36	0,749	0,434	0,774	0,419
7 bis 12	A2	37	0,835	0,372	0,856	0,352
7 bis 12	A2	38	0,884	0,320	0,892	0,311
7 bis 12	A2	39	0,803	0,398	0,825	0,381
7 bis 12	A2	56	0,833	0,373	0,836	0,371
13 bis 18	B1	67	0,641	0,480	0,665	0,472
13 bis 18	B1	72	0,593	0,492	0,607	0,489
13 bis 18	B1	73	0,796	0,403	0,813	0,390
13 bis 18	B1	75	0,714	0,452	0,781	0,414
13 bis 18	B1	76	0,663	0,473	0,667	0,472
13 bis 18	B1	79	0,682	0,466	0,689	0,463
19 bis 24	B2	82	0,692	0,462	0,729	0,445
19 bis 24	B2	85	0,637	0,481	0,622	0,485
19 bis 24	B2	89	0,598	0,491	0,592	0,492
19 bis 24	B2	90	0,469	0,499	0,490	0,500
19 bis 24	B2	91	0,438	0,496	0,495	0,500
19 bis 24	B2	95	0,493	0,500	0,515	0,500
25 bis 30	C1	98	0,532	0,499	0,561	0,497
25 bis 30	C1	100	0,562	0,496	0,575	0,495
25 bis 30	C1	101	0,466	0,499	0,471	0,499
25 bis 30	C1	110	0,367	0,482	0,384	0,487
25 bis 30	C1	111	0,500	0,500	0,498	0,500
25 bis 30	C1	113	0,423	0,494	0,488	0,500
31 bis 36	C2	125	0,439	0,497	0,490	0,500
31 bis 36	C2	134	0,327	0,469	0,352	0,478
31 bis 36	C2	143	0,336	0,473	0,330	0,470
31 bis 36	C2	144	0,317	0,466	0,345	0,476
31 bis 36	C2	153	0,460	0,499	0,414	0,493
31 bis 36	C2	157	0,485	0,500	0,476	0,500

N=804 je Bedingung.

Tabelle A.1-4: Deskriptive Statistik der Punktsummen der zwölf einfachen Items (Schwierigkeitsstufe A) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
ME-Dreieck	-660	1200	932,50	324,92	8,10	1000,00	77,71	1608
ME-Schieberegler	-777	1200	929,70	323,94	8,10	1000,00	77,48	1608
ME-ohne Auszahlungsinformation	-582	1200	914,20	330,30	8,20	1000,00	76,18	1608
ME-mit Auszahlungsinformation	-777	1200	948,10	317,56	7,90	1000,00	79,01	1608
ME-Dreieck-ohne Auszahlungsinformation	-582	1200	921,52	326,18	11,50	1000,00	76,79	804
ME-Dreieck-mit Auszahlungsinformation	-660	1200	943,47	323,48	11,41	1000,00	78,62	804
ME-Schieberegler-ohne Auszahlungsinformation	-460	1200	906,78	334,41	11,79	1000,00	75,57	804
ME-Schieberegler-mit Auszahlungsinformation	-777	1200	952,66	311,66	10,99	1032,00	79,39	804

Tabelle A.1-5: Deskriptive Statistik der Punktsummen der zwölf mittelschwierigen Items (Schwierigkeitsstufe B) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
ME-Dreieck	-892	1200	402,50	458,85	11,40	405,50	33,54	1608
ME-Schieberegler	-1000	1200	416,20	445,34	11,10	400,00	34,68	1608
ME-ohne Auszahlungsinformation	-1000	1200	387,20	456,75	11,40	400,00	32,27	1608
ME-mit Auszahlungsinformation	-905	1200	431,50	446,50	11,10	416,00	35,96	1608
ME-Dreieck-ohne Auszahlungsinformation	-887	1200	382,12	466,34	16,45	400,00	31,84	804
ME-Dreieck-mit Auszahlungsinformation	-892	1200	422,96	450,60	15,89	424,50	35,25	804
ME-Schieberegler-ohne Auszahlungsinformation	-1000	1200	392,36	447,19	15,77	395,50	32,70	804
ME-Schieberegler-mit Auszahlungsinformation	-905	1200	440,02	442,48	15,61	410,50	36,67	804

Tabelle A.1-6: Deskriptive Statistik der Punktsummen der zwölf schwierigen Items (Schwierigkeitsstufe C) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
ME-Dreieck	-1132	1200	52,00	406,06	10,10	0,00	4,33	1608
ME-Schieberegler	-1000	1200	79,90	369,30	9,20	24,00	6,66	1608
ME-ohne Auszahlungsinformation	-1132	1200	54,90	398,56	9,90	0,00	4,58	1608
ME-mit Auszahlungsinformation	-1100	1200	77,00	377,58	9,40	24,00	6,42	1608
ME-Dreieck-ohne Auszahlungsinformation	-1132	1200	54,23	415,54	14,65	0,00	4,52	804
ME-Dreieck-mit Auszahlungsinformation	-1100	1200	49,74	396,60	13,99	0,00	4,14	804
ME-Schieberegler-ohne Auszahlungsinformation	-1000	1200	55,56	381,08	13,44	6,00	4,63	804
ME-Schieberegler-mit Auszahlungsinformation	-800	1200	104,23	355,72	12,55	44,00	8,69	804

Tabelle A.1-7: Deskriptive Statistik der Punktsummen der jeweils zwölf einfachen, mittelschwierigen und schwierigen Items für die Bedingungen mit Multiple-Choice.

Bedingung	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
MC-ohne Auszahlungsinformation	einfach	2	12	10,33	1,98	0,070	11,00	0,861	804
	mittel	1	12	7,41	2,68	0,094	7,00	0,618	804
	schwierig	0	12	5,22	2,45	0,086	5,00	0,435	804
MC-mit Auszahlungsinformation	einfach	2	12	10,55	1,87	0,066	11,00	0,879	804
	mittel	1	12	7,67	2,63	0,093	8,00	0,639	804
	schwierig	0	12	5,38	2,53	0,089	5,00	0,449	804

A.2 *Part-whole-korrigierte* Trennschärfen der Punktauszahlungen

Tabelle A.2-1: *Part-whole-korrigierte* Trennschärfen der Punktauszahlungen der Bedingungen mit Multipler Evaluation und dem Antwortdreieck.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-Dreieck ohne Auszahlungsinformation	ME-Dreieck mit Auszahlungsinformation
1 bis 6	A1	3	0,21	0,23
1 bis 6	A1	4	0,31	0,28
1 bis 6	A1	9	0,27	0,28
1 bis 6	A1	13	0,37	0,41
1 bis 6	A1	15	0,37	0,40
1 bis 6	A1	19	0,24	0,23
7 bis 12	A2	35	0,33	0,31
7 bis 12	A2	36	0,42	0,40
7 bis 12	A2	37	0,27	0,35
7 bis 12	A2	38	0,34	0,33
7 bis 12	A2	39	0,26	0,25
7 bis 12	A2	56	0,37	0,35
13 bis 18	B1	67	0,43	0,39
13 bis 18	B1	72	0,54	0,53
13 bis 18	B1	73	0,26	0,25
13 bis 18	B1	75	0,37	0,38
13 bis 18	B1	76	0,39	0,37
13 bis 18	B1	79	0,36	0,34
19 bis 24	B2	82	0,45	0,37
19 bis 24	B2	85	0,26	0,27
19 bis 24	B2	89	0,26	0,25
19 bis 24	B2	90	0,43	0,38
19 bis 24	B2	91	0,25	0,23
19 bis 24	B2	95	0,38	0,36
25 bis 30	C1	98	0,40	0,38
25 bis 30	C1	100	0,33	0,26
25 bis 30	C1	101	0,32	0,27
25 bis 30	C1	110	0,33	0,28
25 bis 30	C1	111	0,32	0,28
25 bis 30	C1	113	0,35	0,33
31 bis 36	C2	125	0,23	0,22
31 bis 36	C2	134	0,33	0,32
31 bis 36	C2	143	0,27	0,16
31 bis 36	C2	144	0,36	0,33
31 bis 36	C2	153	0,21	0,31
31 bis 36	C2	157	0,35	0,22

N=804 je Bedingung.

Tabelle A.2-2: *Part-whole-korrigierte* Trennschärfen der Punktauszahlungen der Bedingungen mit Multipler Evaluation und den Schiebereglern.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-Schieberegler ohne Auszahlungsinformation	ME-Schieberegler mit Auszahlungsinformation
1 bis 6	A1	3	0,20	0,15
1 bis 6	A1	4	0,30	0,29
1 bis 6	A1	9	0,30	0,34
1 bis 6	A1	13	0,39	0,44
1 bis 6	A1	15	0,43	0,44
1 bis 6	A1	19	0,26	0,27
7 bis 12	A2	35	0,43	0,37
7 bis 12	A2	36	0,43	0,44
7 bis 12	A2	37	0,30	0,24
7 bis 12	A2	38	0,33	0,37
7 bis 12	A2	39	0,30	0,24
7 bis 12	A2	56	0,36	0,38
13 bis 18	B1	67	0,41	0,42
13 bis 18	B1	72	0,55	0,53
13 bis 18	B1	73	0,21	0,26
13 bis 18	B1	75	0,39	0,44
13 bis 18	B1	76	0,40	0,44
13 bis 18	B1	79	0,40	0,35
19 bis 24	B2	82	0,37	0,39
19 bis 24	B2	85	0,25	0,31
19 bis 24	B2	89	0,23	0,25
19 bis 24	B2	90	0,44	0,47
19 bis 24	B2	91	0,22	0,25
19 bis 24	B2	95	0,34	0,37
25 bis 30	C1	98	0,34	0,39
25 bis 30	C1	100	0,30	0,36
25 bis 30	C1	101	0,31	0,24
25 bis 30	C1	110	0,25	0,23
25 bis 30	C1	111	0,30	0,39
25 bis 30	C1	113	0,34	0,33
31 bis 36	C2	125	0,19	0,24
31 bis 36	C2	134	0,27	0,23
31 bis 36	C2	143	0,18	0,23
31 bis 36	C2	144	0,32	0,31
31 bis 36	C2	153	0,25	0,29
31 bis 36	C2	157	0,30	0,33

N=804 je Bedingung.

Tabelle A.2-3: *Part-whole-korrigierte* Trennschärfen der Punktauszahlungen der Bedingungen mit Multiple-Choice.

Position im Test	GER-Stufe	Item Datenbank Nr.	MC ohne Auszahlungsinformation	MC mit Auszahlungsinformation
1 bis 6	A1	3	0,15	0,17
1 bis 6	A1	4	0,33	0,29
1 bis 6	A1	9	0,33	0,34
1 bis 6	A1	13	0,41	0,42
1 bis 6	A1	15	0,38	0,34
1 bis 6	A1	19	0,24	0,21
7 bis 12	A2	35	0,36	0,30
7 bis 12	A2	36	0,40	0,40
7 bis 12	A2	37	0,25	0,26
7 bis 12	A2	38	0,32	0,34
7 bis 12	A2	39	0,26	0,26
7 bis 12	A2	56	0,35	0,39
13 bis 18	B1	67	0,45	0,44
13 bis 18	B1	72	0,52	0,53
13 bis 18	B1	73	0,17	0,23
13 bis 18	B1	75	0,40	0,31
13 bis 18	B1	76	0,45	0,39
13 bis 18	B1	79	0,40	0,36
19 bis 24	B2	82	0,38	0,30
19 bis 24	B2	85	0,25	0,30
19 bis 24	B2	89	0,28	0,23
19 bis 24	B2	90	0,36	0,49
19 bis 24	B2	91	0,20	0,22
19 bis 24	B2	95	0,32	0,28
25 bis 30	C1	98	0,37	0,41
25 bis 30	C1	100	0,17	0,25
25 bis 30	C1	101	0,26	0,29
25 bis 30	C1	110	0,33	0,34
25 bis 30	C1	111	0,31	0,23
25 bis 30	C1	113	0,31	0,35
31 bis 36	C2	125	0,22	0,12
31 bis 36	C2	134	0,34	0,34
31 bis 36	C2	143	0,06	0,11
31 bis 36	C2	144	0,32	0,41
31 bis 36	C2	153	0,15	0,21
31 bis 36	C2	157	0,21	0,24

N=804 je Bedingung.

A.3 Deskriptive Statistik der Punktsummen nach einer Realismuskorrektur

Tabelle A.3-1: Deskriptive Statistik der Punktsummen der realismuskorrigierten Punktauszahlungen aller 36 Items.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
ME-Dreieck-ohne Auszahlungsinformation	-14	3600	1223,48	881,90	31,10	1069,50	33,99	804
ME-Dreieck-mit Auszahlungsinformation	0	3600	1251,42	822,34	29,00	1156,00	34,76	804
ME-Schieberegler-ohne Auszahlungsinformation	0	3600	1216,49	843,08	29,73	1104,50	33,79	804
ME-Schieberegler-mit Auszahlungsinformation	0	3600	1338,14	843,08	29,73	1104,50	37,17	804

Tabelle A.3-2: Deskriptive Statistik der Punktsummen der realismuskorrigierten Punktauszahlungen der zwölf einfachen Items.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
ME-Dreieck-ohne Auszahlungsinformation	-52,00	1200	691,38	331,20	11,68	725,00	57,61	804
ME-Dreieck-mit Auszahlungsinformation	-88,00	1200	709,00	310,56	10,95	766,00	59,08	804
ME-Schieberegler-ohne Auszahlungsinformation	-28,00	1200	694,00	330,63	11,66	745,50	57,83	804
ME-Schieberegler-mit Auszahlungsinformation	-18,00	1200	758,66	311,28	10,98	819,50	63,22	804

Tabelle A.3-3: Deskriptive Statistik der Punktsummen der realismuskorrigierten Punktauszahlungen der zwölf mittelschwierigen Items.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
ME-Dreieck-ohne Auszahlungsinformation	-167,00	1200	385,58	351,04	12,38	317,00	32,13	804
ME-Dreieck-mit Auszahlungsinformation	-125,00	1200	405,68	339,88	11,99	344,50	33,81	804
ME-Schieberegler-ohne Auszahlungsinformation	-177,00	1200	389,03	345,23	12,18	320,00	32,42	804
ME-Schieberegler-mit Auszahlungsinformation	-198,00	1200	425,26	353,61	12,47	356,00	35,44	804

Tabelle A.3-4: Deskriptive Statistik der Punktsummen der realismuskorrigierten Punktauszahlungen der zwölf schwierigen Items.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	N
ME-Dreieck-ohne Auszahlungsinformation	-364	1200	146,52	300,76	10,61	30,00	12,21	804
ME-Dreieck-mit Auszahlungsinformation	-302	1200	136,74	280,66	9,90	40,00	11,39	804
ME-Schieberegler-ohne Auszahlungsinformation	-340	1200	133,46	280,66	9,69	32,50	11,12	804
ME-Schieberegler-mit Auszahlungsinformation	-284	1200	154,21	278,41	9,82	56,50	12,85	804

B. Daten des zweiten Experiments

B.1 Punktauszahlungen

Tabelle B.1-1: Mittelwert und Standardabweichung je Item für die Bedingungen „ME-linear 0 bis 100 Punkte“ und „ME logarithmisch -50 bis 100 Punkte“.

			ME-linear		ME-logarithmisch	
Position	GER-	Item	0 bis 100 Punkte		-50 bis 100 Punkte	
im Test	Stufe	Datenbank Nr.	Mittelwert	Standardabweichung	Mittelwert	Standardabweichung
1 bis 6	A1	3	97,4	14,5	96,1	22,8
1 bis 6	A1	4	94,3	20,7	93,3	27,4
1 bis 6	A1	9	90,5	26,7	88,8	34,5
1 bis 6	A1	13	83,4	35,3	78,6	49,2
1 bis 6	A1	15	86,9	30,5	81,2	43,5
1 bis 6	A1	19	94,1	20,2	91,6	28,4
7 bis 12	A2	35	80,8	36,5	76,1	49,7
7 bis 12	A2	36	76,0	39,7	65,9	56,7
7 bis 12	A2	37	83,3	34,0	78,6	44,3
7 bis 12	A2	38	87,0	30,3	84,0	39,5
7 bis 12	A2	39	78,8	36,5	70,8	52,9
7 bis 12	A2	56	84,9	31,9	77,1	46,3
13 bis 18	B1	67	67,1	41,4	52,8	60,3
13 bis 18	B1	72	61,4	45,7	48,6	64,8
13 bis 18	B1	73	80,1	35,7	69,0	53,6
13 bis 18	B1	75	74,0	37,3	64,8	53,5
13 bis 18	B1	76	66,0	43,3	54,0	61,1
13 bis 18	B1	79	69,3	41,7	50,7	63,5
19 bis 24	B2	82	67,5	40,8	58,4	56,5
19 bis 24	B2	85	61,3	42,3	40,8	61,2
19 bis 24	B2	89	58,7	43,7	34,5	64,4
19 bis 24	B2	90	54,2	41,4	29,2	61,9
19 bis 24	B2	91	42,8	43,0	21,1	64,4
19 bis 24	B2	95	50,6	41,4	28,0	61,4
25 bis 30	C1	98	56,1	39,3	34,6	56,9
25 bis 30	C1	100	52,3	36,6	28,8	52,2
25 bis 30	C1	101	48,6	42,5	23,4	60,0
25 bis 30	C1	110	42,3	37,4	13,4	54,5
25 bis 30	C1	111	50,1	39,0	27,1	56,8
25 bis 30	C1	113	46,0	41,1	21,3	59,2
31 bis 36	C2	125	44,8	30,4	17,6	43,7
31 bis 36	C2	134	37,0	34,8	7,0	50,4
31 bis 36	C2	143	36,9	28,2	7,6	41,8
31 bis 36	C2	144	39,9	35,8	8,7	51,8
31 bis 36	C2	153	40,8	37,3	12,9	53,1
31 bis 36	C2	157	44,6	37,0	18,3	54,5

N=808 je Bedingung.

Tabelle B.1-2: Mittelwert und Standardabweichung je Item für die Bedingungen „ME logarithmisch -100 bis 100 Punkte“ und „ME logarithmisch -300 bis 100 Punkte“.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-logarithmisch -100 bis 100 Punkte Mittelwert	ME-logarithmisch -100 bis 100 Punkte Standardabweichung	ME-logarithmisch -300 bis 100 Punkte Mittelwert	ME-logarithmisch -300 bis 100 Punkte Standardabweichung
1 bis 6	A1	3	95,9	25,5	94,7	42,3
1 bis 6	A1	4	91,3	36,0	84,8	68,7
1 bis 6	A1	9	85,3	46,0	73,7	88,2
1 bis 6	A1	13	75,6	60,9	46,9	126,2
1 bis 6	A1	15	76,1	56,9	68,2	91,5
1 bis 6	A1	19	90,5	33,9	83,6	68,9
7 bis 12	A2	35	69,7	64,2	53,6	112,6
7 bis 12	A2	36	57,8	73,9	35,4	129,4
7 bis 12	A2	37	70,2	62,5	60,7	98,9
7 bis 12	A2	38	79,8	50,6	70,1	90,0
7 bis 12	A2	39	66,7	63,6	33,9	130,8
7 bis 12	A2	56	75,6	55,8	64,1	94,4
13 bis 18	B1	67	47,9	73,4	19,1	133,1
13 bis 18	B1	72	35,4	83,9	-10,0	159,1
13 bis 18	B1	73	65,0	65,0	45,1	117,0
13 bis 18	B1	75	57,2	66,2	45,0	105,1
13 bis 18	B1	76	37,9	81,0	10,7	143,3
13 bis 18	B1	79	44,1	76,4	7,0	147,5
19 bis 24	B2	82	47,6	71,1	26,9	122,5
19 bis 24	B2	85	30,6	77,8	-5,7	145,9
19 bis 24	B2	89	27,3	82,2	-13,9	152,7
19 bis 24	B2	90	17,8	76,0	-16,6	140,8
19 bis 24	B2	91	3,7	82,8	-53,6	161,6
19 bis 24	B2	95	14,3	78,3	-18,1	142,3
25 bis 30	C1	98	26,9	72,8	2,4	125,6
25 bis 30	C1	100	24,2	62,4	9,0	103,5
25 bis 30	C1	101	10,5	79,0	-28,9	146,6
25 bis 30	C1	110	2,6	72,2	-32,4	136,4
25 bis 30	C1	111	17,6	69,1	-9,8	127,4
25 bis 30	C1	113	6,9	75,4	-28,7	139,5
31 bis 36	C2	125	10,4	57,0	-7,9	103,7
31 bis 36	C2	134	-5,9	65,2	-31,9	127,0
31 bis 36	C2	143	-1,9	52,2	-24,1	106,3
31 bis 36	C2	144	-3,2	65,0	-31,6	123,9
31 bis 36	C2	153	1,2	68,6	-34,3	131,7
31 bis 36	C2	157	11,0	66,8	-19,9	125,0

N=808 je Bedingung.

Tabelle B.1-3: Mittelwert und Standardabweichung der Punktauszahlung je Item für die Bedingungen mit Multiple-Choice.

Position im Test	GER-Stufe	Item Datenbank Nr.	MC-dichotom 0 oder 1 Punkt Mittelwert	Standardabweichung
1 bis 6	A1	3	0,960	0,195
1 bis 6	A1	4	0,917	0,276
1 bis 6	A1	9	0,912	0,283
1 bis 6	A1	13	0,828	0,378
1 bis 6	A1	15	0,887	0,316
1 bis 6	A1	19	0,939	0,239
7 bis 12	A2	35	0,791	0,407
7 bis 12	A2	36	0,730	0,444
7 bis 12	A2	37	0,849	0,358
7 bis 12	A2	38	0,896	0,305
7 bis 12	A2	39	0,808	0,394
7 bis 12	A2	56	0,848	0,359
13 bis 18	B1	67	0,666	0,472
13 bis 18	B1	72	0,610	0,488
13 bis 18	B1	73	0,819	0,385
13 bis 18	B1	75	0,778	0,416
13 bis 18	B1	76	0,651	0,477
13 bis 18	B1	79	0,645	0,479
19 bis 24	B2	82	0,719	0,450
19 bis 24	B2	85	0,604	0,489
19 bis 24	B2	89	0,609	0,488
19 bis 24	B2	90	0,509	0,500
19 bis 24	B2	91	0,457	0,498
19 bis 24	B2	95	0,496	0,500
25 bis 30	C1	98	0,582	0,494
25 bis 30	C1	100	0,593	0,492
25 bis 30	C1	101	0,484	0,500
25 bis 30	C1	110	0,416	0,493
25 bis 30	C1	111	0,527	0,500
25 bis 30	C1	113	0,490	0,500
31 bis 36	C2	125	0,450	0,498
31 bis 36	C2	134	0,337	0,473
31 bis 36	C2	143	0,342	0,475
31 bis 36	C2	144	0,340	0,474
31 bis 36	C2	153	0,385	0,487
31 bis 36	C2	157	0,434	0,496

N=808 je Bedingung.

Tabelle B.1-4: Deskriptive Statistik der Punktsummen der einfachen Items (Schwierigkeitsstufe A) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
ME-linear 0 bis 100 Punkte	279	1200	1037,43	196,23	6,90	1100,00	86,45	808
ME-logarithmisch -50 bis 100 Punkte	-21	1200	982,15	263,18	9,26	1050,00	81,85	808
ME-logarithmisch -100 bis 100 Punkte	-413	1200	934,52	328,06	11,54	1000,00	77,88	808
ME-logarithmisch -300 bis 100 Punkte	-2244	1200	769,69	530,29	18,66	971,00	64,14	808

Tabelle B.1-5: Deskriptive Statistik der Punktsummen für mittelschwierige Items (Schwierigkeitsstufe B) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
ME-linear 0 bis 100 Punkte	191	1200	753,05	265,90	9,35	748,00	62,75	808
ME-logarithmisch -50 bis 100 Punkte	-312	1200	551,85	364,01	12,81	529,50	45,99	808
ME-logarithmisch -100 bis 100 Punkte	-1000	1200	428,77	453,41	16,60	400,00	35,73	808
ME-logarithmisch -300 bis 100 Punkte	-2400	1200	35,75	738,59	25,98	18,00	2,98	808

Tabelle B.1-6: Deskriptive Statistik der Punktsummen für schwierige Items (Schwierigkeitsstufe C) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
ME-linear 0 bis 100 Punkte	0	1200	538,00	253,12	8,90	500,00	14,94	808
ME-logarithmisch -50 bis 100 Punkte	-450	1200	220,62	328,95	11,57	125,00	6,13	808
ME-logarithmisch -100 bis 100 Punkte	-1200	1200	100,32	370,85	13,05	36,50	2,79	808
ME-logarithmisch -300 bis 100 Punkte	-3600	1200	-237,95	701,32	24,61	-78,50	-6,61	808

Tabelle B.1-7: Deskriptive Statistik der Punktsummen der einfachen, mittelschwierigen und schwierigen Items für die Bedingung mit Multiple-Choice.

GER-Stufe	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/Item	N
A	einfach	1	12	10,37	2,09	0,074	11,00	0,864	808
B	mittel	0	12	7,56	2,65	0,093	8,00	0,630	808
C	schwierig	0	12	5,38	2,53	0,089	5,00	0,447	808

B.2 *Part-whole-korrigierte* Trennschärfen der Punktauszahlungen

Tabelle B.2-1: *Part-whole-korrigierte* Trennschärfen der Items, berechnet über ausgezahlte Punkte für die Bedingungen „ME-linear 0 bis 100 Punkte“ und „ME-logarithmisch -50 bis 100 Punkte“.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-linear 0 bis 100 Punkte	ME-logarithmisch -50 bis 100 Punkte
1 bis 6	A1	3	0,23	0,11
1 bis 6	A1	4	0,36	0,28
1 bis 6	A1	9	0,38	0,31
1 bis 6	A1	13	0,47	0,43
1 bis 6	A1	15	0,44	0,43
1 bis 6	A1	19	0,30	0,30
7 bis 12	A2	35	0,47	0,41
7 bis 12	A2	36	0,51	0,50
7 bis 12	A2	37	0,34	0,33
7 bis 12	A2	38	0,39	0,42
7 bis 12	A2	39	0,29	0,27
7 bis 12	A2	56	0,43	0,43
13 bis 18	B1	67	0,50	0,43
13 bis 18	B1	72	0,59	0,55
13 bis 18	B1	73	0,29	0,30
13 bis 18	B1	75	0,49	0,37
13 bis 18	B1	76	0,38	0,37
13 bis 18	B1	79	0,39	0,39
19 bis 24	B2	82	0,46	0,40
19 bis 24	B2	85	0,41	0,36
19 bis 24	B2	89	0,41	0,26
19 bis 24	B2	90	0,53	0,55
19 bis 24	B2	91	0,30	0,32
19 bis 24	B2	95	0,44	0,45
25 bis 30	C1	98	0,48	0,45
25 bis 30	C1	100	0,40	0,43
25 bis 30	C1	101	0,35	0,36
25 bis 30	C1	110	0,37	0,38
25 bis 30	C1	111	0,42	0,44
25 bis 30	C1	113	0,48	0,44
31 bis 36	C2	125	0,31	0,28
31 bis 36	C2	134	0,33	0,36
31 bis 36	C2	143	0,29	0,32
31 bis 36	C2	144	0,45	0,41
31 bis 36	C2	153	0,30	0,30
31 bis 36	C2	157	0,33	0,36

N=808 je Bedingung.

Tabelle B.2-2: *Part-whole-korrigierte* Trennschärfen der Items, berechnet über ausgezahlte Punkte für die Bedingungen „ME logarithmisch -100 bis 100 Punkte“ und „ME logarithmisch -300 bis 100 Punkte“.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-logarithmisch -100 bis 100 Punkte	ME-logarithmisch -300 bis 100 Punkte
1 bis 6	A1	3	0,18	0,13
1 bis 6	A1	4	0,28	0,17
1 bis 6	A1	9	0,32	0,25
1 bis 6	A1	13	0,39	0,33
1 bis 6	A1	15	0,40	0,32
1 bis 6	A1	19	0,31	0,19
7 bis 12	A2	35	0,34	0,36
7 bis 12	A2	36	0,47	0,33
7 bis 12	A2	37	0,34	0,23
7 bis 12	A2	38	0,38	0,26
7 bis 12	A2	39	0,28	0,20
7 bis 12	A2	56	0,37	0,30
13 bis 18	B1	67	0,40	0,31
13 bis 18	B1	72	0,58	0,44
13 bis 18	B1	73	0,26	0,18
13 bis 18	B1	75	0,38	0,27
13 bis 18	B1	76	0,40	0,32
13 bis 18	B1	79	0,36	0,31
19 bis 24	B2	82	0,35	0,34
19 bis 24	B2	85	0,32	0,29
19 bis 24	B2	89	0,31	0,22
19 bis 24	B2	90	0,5	0,38
19 bis 24	B2	91	0,31	0,23
19 bis 24	B2	95	0,39	0,29
25 bis 30	C1	98	0,43	0,32
25 bis 30	C1	100	0,32	0,32
25 bis 30	C1	101	0,28	0,34
25 bis 30	C1	110	0,29	0,29
25 bis 30	C1	111	0,40	0,25
25 bis 30	C1	113	0,39	0,33
31 bis 36	C2	125	0,22	0,16
31 bis 36	C2	134	0,21	0,25
31 bis 36	C2	143	0,20	0,23
31 bis 36	C2	144	0,31	0,39
31 bis 36	C2	153	0,23	0,31
31 bis 36	C2	157	0,30	0,31

N=808 je Bedingung.

Tabelle B.2-3: *Part-whole-korrigierte* Trennschärfen der Items, berechnet über ausgezahlte Punkte für die Bedingungen mit Multiple-Choice.

Position im Test	GER-Stufe	Item Datenbank Nr.	MC- 0 oder 1 Punkt
1 bis 6	A1	3	0,25
1 bis 6	A1	4	0,38
1 bis 6	A1	9	0,34
1 bis 6	A1	13	0,42
1 bis 6	A1	15	0,42
1 bis 6	A1	19	0,33
7 bis 12	A2	35	0,37
7 bis 12	A2	36	0,49
7 bis 12	A2	37	0,23
7 bis 12	A2	38	0,36
7 bis 12	A2	39	0,28
7 bis 12	A2	56	0,40
13 bis 18	B1	67	0,39
13 bis 18	B1	72	0,57
13 bis 18	B1	73	0,23
13 bis 18	B1	75	0,34
13 bis 18	B1	76	0,42
13 bis 18	B1	79	0,35
19 bis 24	B2	82	0,41
19 bis 24	B2	85	0,24
19 bis 24	B2	89	0,27
19 bis 24	B2	90	0,45
19 bis 24	B2	91	0,24
19 bis 24	B2	95	0,31
25 bis 30	C1	98	0,33
25 bis 30	C1	100	0,23
25 bis 30	C1	101	0,33
25 bis 30	C1	110	0,29
25 bis 30	C1	111	0,29
25 bis 30	C1	113	0,36
31 bis 36	C2	125	0,22
31 bis 36	C2	134	0,26
31 bis 36	C2	143	0,13
31 bis 36	C2	144	0,38
31 bis 36	C2	153	0,22
31 bis 36	C2	157	0,24

N=808 je Bedingung.

B.3 Deskriptive Statistik der Punktsummen nach einer Realismuskorrektur

Tabelle B.3-1: Realismuskorrigierte Punktsummen über alle Schwierigkeitsstufen.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte	1197	3600	2012,99	607,84	21,38	1903,50	55,92	64,7
ME-logarithmisch -50 bis 100 Punkte	5	3600	1292,67	852,85	30,00	1156,50	35,91	48,7
ME-logarithmisch -100 bis 100 Punkte	-1	3600	1305,55	870,36	30,62	1179,00	36,27	40,7
ME-logarithmisch -300 bis 100 Punkte	-56	3600	1243,00	860,18	30,26	1108,00	34,53	15,8

N=808 je Bedingung, Anzahl der Items=36.

Tabelle B.3-2: Realismuskorrigierte Punktsummen der einfachen Items (GER-Stufe A).

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte	393	1200	823,04	224,91	7,91	859,00	68,59	86,5
ME-logarithmisch -50 bis 100 Punkte	-46	1200	682,46	306,78	10,79	720,00	56,87	81,8
ME-logarithmisch -100 bis 100 Punkte	-38	1200	733,94	325,79	11,46	805,50	61,16	77,9
ME-logarithmisch -300 bis 100 Punkte	-233	1200	756,85	336,16	11,83	818,00	63,07	64,1

N=808 je Bedingung, Anzahl der Items=12.

Tabelle B.3-3: Realismuskorrigierte Punktsummen der mittelschwierigen Items (GER-Stufe B).

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte	323	1200	666,64	237,59	8,36	609,00	55,55	62,8
ME-logarithmisch -50 bis 100 Punkte	-91	1200	420,11	335,37	11,80	352,00	35,01	46,0
ME-logarithmisch -100 bis 100 Punkte	-204	1200	416,04	360,80	12,69	348,00	34,67	35,7
ME-logarithmisch -300 bis 100 Punkte	-250	1200	371,47	380,69	13,39	264,00	31,96	3,0

N=808 je Bedingung, Anzahl der Items=12.

Tabelle B.3-4: Realismuskorrigierte Punktsummen der schwierigen Items (GER-Stufe C).

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte	268	1200	523,30	193,07	6,79	438,50	43,61	14,9
ME-logarithmisch -50 bis 100 Punkte	-200	1200	190,11	282,65	9,94	73,50	15,84	6,1
ME-logarithmisch -100 bis 100 Punkte	-276	1200	155,58	286,87	10,09	52,50	12,96	2,8
ME-logarithmisch -300 bis 100 Punkte	-565	1200	114,68	296,55	10,43	15,50	9,56	-6,6

N=808 je Bedingung, Anzahl der Items=12.

C. Daten des dritten Experiments

C.1 Punktauszahlungen

Tabelle C.1-1: Mittelwert und Standardabweichung der Punktauszahlung je Item für die Bedingungen „ME-linear 0 bis 100 Punkte-2 Antwortoptionen“ und „ME-linear 0 bis 100 Punkte-3 Antwortoptionen“.

			ME-linear 0 bis 100 Punkte 2 Antwortoptionen		ME-linear 0 bis 100 Punkte 3 Antwortoptionen	
Position im Test	GER-Stufe	Item Datenbank Nr.	Mittelwert	Standardabweichung	Mittelwert	Standardabweichung
1 bis 6	A1	3	98,0	13,6	98,3	11,8
1 bis 6	A1	4	95,2	19,4	96,4	16,2
1 bis 6	A1	9	94,1	20,9	92,6	23,4
1 bis 6	A1	13	87,4	30,8	87,1	31,4
1 bis 6	A1	15	86,5	30,0	90,0	26,4
1 bis 6	A1	19	95,6	17,3	95,4	17,1
7 bis 12	A2	35	88,9	28,4	85,4	32,0
7 bis 12	A2	36	79,2	37,3	80,6	35,9
7 bis 12	A2	37	84,3	32,2	87,4	28,5
7 bis 12	A2	38	90,4	24,8	90,7	25,8
7 bis 12	A2	39	84,0	32,7	81,9	33,6
7 bis 12	A2	56	86,0	30,2	88,1	28,8
13 bis 18	B1	67	75,5	36,4	72,3	40,1
13 bis 18	B1	72	70,6	42,2	67,6	43,7
13 bis 18	B1	73	83,2	32,6	82,3	33,9
13 bis 18	B1	75	80,5	32,8	78,3	35,3
13 bis 18	B1	76	78,5	35,9	73,8	39,8
13 bis 18	B1	79	76,9	37,0	71,7	41,3
19 bis 24	B2	82	82,1	32,3	74,8	38,0
19 bis 24	B2	85	70,2	38,8	65,0	40,9
19 bis 24	B2	89	67,0	40,8	61,6	42,9
19 bis 24	B2	90	67,0	39,7	55,8	42,5
19 bis 24	B2	91	64,6	41,5	48,4	43,5
19 bis 24	B2	95	59,7	39,4	51,9	42,5
25 bis 30	C1	98	70,7	38,5	61,5	38,3
25 bis 30	C1	100	68,5	32,7	55,0	36,4
25 bis 30	C1	101	66,1	38,5	54,0	41,8
25 bis 30	C1	110	51,8	38,7	41,8	38,2
25 bis 30	C1	111	65,2	36,8	52,9	39,3
25 bis 30	C1	113	62,2	40,5	53,9	41,1
31 bis 36	C2	125	53,4	30,2	45,4	31,1
31 bis 36	C2	134	51,6	35,7	37,5	35,1
31 bis 36	C2	143	53,8	30,4	37,0	29,9
31 bis 36	C2	144	52,1	37,0	41,2	36,4
31 bis 36	C2	153	57,1	38,2	41,6	37,6
31 bis 36	C2	157	59,3	37,3	49,0	37,7

N=1100 je Bedingung.

Tabelle C.1-2: Mittelwert und Standardabweichung der Punktauszahlung je Item für die Bedingungen „ME -logarithmisch -300 bis 100 Punkte-2 Antwortoptionen“ und „ME- logarithmisch -300 bis 100 Punkte-3 Antwortoptionen“.

			ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen		ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	
Position im Test	GER-Stufe	Item Datenbank Nr.	Mittelwert	Standardabweichung	Mittelwert	Standardabweichung
1 bis 6	A1	3	95,3	40,5	94,1	45,1
1 bis 6	A1	4	90,9	52,7	92,8	45,9
1 bis 6	A1	9	84,8	63,5	81,4	72,0
1 bis 6	A1	13	60,5	111,4	63,4	104,7
1 bis 6	A1	15	65,6	97,2	75,8	80,7
1 bis 6	A1	19	86,5	64,4	93,3	33,7
7 bis 12	A2	35	71,4	87,3	62,6	99,6
7 bis 12	A2	36	45,7	116,9	39,1	130,5
7 bis 12	A2	37	60,3	97,3	63,1	98,2
7 bis 12	A2	38	74,7	79,0	77,8	75,5
7 bis 12	A2	39	57,1	102,9	47,2	117,4
7 bis 12	A2	56	67,5	87,1	69,9	88,2
13 bis 18	B1	67	27,3	119,4	20,9	131,5
13 bis 18	B1	72	22,6	135,9	9,6	148,1
13 bis 18	B1	73	54,0	98,3	51,5	108,6
13 bis 18	B1	75	44,2	102,4	46,6	104,9
13 bis 18	B1	76	30,5	125,6	21,5	136,2
13 bis 18	B1	79	27,8	130,5	22,9	136,2
19 bis 24	B2	82	50,2	102,9	35,3	116,7
19 bis 24	B2	85	28,6	116,8	9,7	135,9
19 bis 24	B2	89	12,4	135,0	-19,4	154,2
19 bis 24	B2	90	9,0	127,4	-4,5	131,1
19 bis 24	B2	91	-2,9	137,2	-43,3	157,8
19 bis 24	B2	95	-11,0	128,0	-13,5	137,5
25 bis 30	C1	98	32,4	109,4	11,6	116,9
25 bis 30	C1	100	24,6	92,9	8,7	104,9
25 bis 30	C1	101	18,2	115,1	-15,3	139,5
25 bis 30	C1	110	-30,0	133,2	-36,6	137,2
25 bis 30	C1	111	12,2	112,2	-5,0	122,6
25 bis 30	C1	113	0,8	129,0	-24,3	140,5
31 bis 36	C2	125	-4,7	92,0	-6,3	102,8
31 bis 36	C2	134	-25,2	121,9	-30,3	126,3
31 bis 36	C2	143	-12,7	103,0	-17,9	101,7
31 bis 36	C2	144	-24,9	122,9	-30,4	125,8
31 bis 36	C2	153	-8,7	122,3	-30,2	127,8
31 bis 36	C2	157	-6,0	118,6	-17,1	121,3

N=1100 je Bedingung.

Tabelle C.1-3: Mittelwert und Standardabweichung der Punktauszahlung je Item für die Bedingungen mit Multiple-Choice „MC- 0 oder 1 Punkt-2 Antwortoptionen“ und „MC- 0 oder 1 Punkt-3 Antwortoptionen“.

			MC- 0 oder 1 Punkt 2 Antwortoptionen		MC- 0 oder 1 Punkt 3 Antwortoptionen	
Position im Test	GER-Stufe	Item Datenbank Nr.	Mittelwert	Standardabweichung	Mittelwert	Standardabweichung
1 bis 6	A1	3	0,985	0,123	0,981	0,137
1 bis 6	A1	4	0,965	0,185	0,974	0,160
1 bis 6	A1	9	0,951	0,216	0,932	0,252
1 bis 6	A1	13	0,873	0,333	0,872	0,334
1 bis 6	A1	15	0,909	0,288	0,911	0,285
1 bis 6	A1	19	0,967	0,178	0,965	0,185
7 bis 12	A2	35	0,886	0,318	0,837	0,369
7 bis 12	A2	36	0,800	0,400	0,781	0,414
7 bis 12	A2	37	0,861	0,346	0,856	0,351
7 bis 12	A2	38	0,912	0,284	0,920	0,271
7 bis 12	A2	39	0,857	0,350	0,847	0,360
7 bis 12	A2	56	0,883	0,322	0,874	0,332
13 bis 18	B1	67	0,783	0,413	0,735	0,442
13 bis 18	B1	72	0,706	0,456	0,668	0,471
13 bis 18	B1	73	0,835	0,371	0,810	0,392
13 bis 18	B1	75	0,810	0,392	0,805	0,396
13 bis 18	B1	76	0,783	0,413	0,717	0,451
13 bis 18	B1	79	0,772	0,420	0,714	0,452
19 bis 24	B2	82	0,847	0,360	0,745	0,436
19 bis 24	B2	85	0,721	0,449	0,657	0,475
19 bis 24	B2	89	0,719	0,450	0,617	0,486
19 bis 24	B2	90	0,652	0,477	0,565	0,496
19 bis 24	B2	91	0,675	0,468	0,468	0,499
19 bis 24	B2	95	0,602	0,490	0,537	0,499
25 bis 30	C1	98	0,712	0,453	0,607	0,489
25 bis 30	C1	100	0,730	0,444	0,607	0,489
25 bis 30	C1	101	0,671	0,470	0,542	0,498
25 bis 30	C1	110	0,502	0,500	0,396	0,489
25 bis 30	C1	111	0,685	0,465	0,544	0,498
25 bis 30	C1	113	0,632	0,483	0,528	0,499
31 bis 36	C2	125	0,524	0,500	0,461	0,499
31 bis 36	C2	134	0,483	0,500	0,391	0,488
31 bis 36	C2	143	0,529	0,499	0,349	0,477
31 bis 36	C2	144	0,483	0,500	0,404	0,491
31 bis 36	C2	153	0,577	0,494	0,435	0,496
31 bis 36	C2	157	0,575	0,495	0,494	0,500

N=1100 je Bedingung.

Tabelle C.1-4: Deskriptive Statistik der Punktsummen für einfache Items (GER-Stufe A) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardfehler	Standardabweichung	Median	Mittelwert/ Item	Anzahl Items	N
ME-2 Antwortoptionen	-1972	1200	964,96	7,75	363,31	1100,00	80,41	12	2200
ME-3 Antwortoptionen	-2295	1200	967,30	8,03	376,74	1114,00	80,61	12	2200
ME-linear 0 bis 100 Punkte	331	1200	1071,75	3,37	158,01	1100,00	89,31	12	2200
ME-logarithmisch -300 bis 100 Punkte	-2295	1200	860,51	10,15	476,06	1100,00	71,71	12	2200
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	406	1200	1069,53	4,63	153,63	1100,00	89,13	12	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	331	1200	1073,98	4,89	162,32	1120,00	89,50	12	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-1972	1200	860,39	14,10	467,56	1100,00	71,70	12	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-2295	1200	860,62	14,61	484,63	1102,00	71,72	12	1100

Tabelle C.1-5: Deskriptive Statistik der Punktsummen für mittelschwierige Items (GER-Stufe B) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardfehler	Standardabweichung	Median	Mittelwert/ Item	Anzahl Items	N
ME-2 Antwortoptionen	-2580	1200	584,16	11,30	530,17	735,00	48,68	12	2200
ME-3 Antwortoptionen	-2800	1200	470,53	13,54	634,99	616,00	39,21	12	2200
ME-linear 0 bis 100 Punkte	150	1200	839,70	4,89	229,15	854,50	69,98	12	2200
ME-logarithmisch -300 bis 100 Punkte	-2800	1200	214,99	14,19	665,60	262,00	17,92	12	2200
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	300	1200	875,72	5,89	195,47	880,00	72,98	12	1100
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	150	1200	803,69	7,64	253,50	800,50	66,97	12	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-2580	1200	292,60	17,94	595,01	310,00	24,38	12	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-2800	1200	137,37	21,75	721,35	212,50	11,45	12	1100

Tabelle C.1-6: Deskriptive Statistik der Punktsummen für schwierige Items (GER-Stufe C) für die Bedingungen mit Multipler Evaluation.

Bedingung	Minimum	Maximum	Mittelwert	Standardfehler	Standardabweichung	Median	Mittelwert/ Item	Anzahl Items	N
ME-2 Antwortoptionen	-2500	1200	343,85	12,44	583,68	520,00	28,65	12	2200
ME-3 Antwortoptionen	-3200	1200	188,82	13,55	635,66	369,50	15,74	12	2200
ME-linear 0 bis 100 Punkte	100	1200	641,30	4,80	225,10	600,50	53,44	12	2200
ME-logarithmisch -300 bis 100 Punkte	-3200	1200	-108,62	13,90	651,75	-25,50	-9,05	12	2200
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	150	1200	711,75	6,17	204,73	670,00	59,31	12	1100
ME-linear 0 bis 100 Punkte- 3 Antwortoptionen	100	1200	570,85	6,71	222,56	515,00	47,57	12	1100
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-2500	1200	-24,05	18,31	607,31	22,50	-2,00	12	1100
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-3200	1200	-193,20	20,60	683,23	-64,00	-16,10	12	1100

Tabelle C.1-7: Deskriptive Statistik der Punktsummen für die Bedingungen mit Multiple-Choice für alle 36 Items des Tests und für die jeweils zwölf einfachen, mittelschwierigen und schwierigen Items.

Bedingung	Schwierigkeit	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Anzahl Items
MC 2 Antwortoptionen	einfach (A)	4	12	10,85	1,59	0,045	11,00	0,904	12
	mittel (B)	3	12	8,91	20,06	0,062	9,00	0,742	12
	schwierig (C)	1	12	7,10	2,40	0,100	7,00	0,592	12
	alle	12	36	26,85	4,88	0,147	27,00	0,746	36
MC 3 Antwortoptionen	einfach (A)	2	12	10,75	1,67	0,050	11,00	0,896	12
	mittel (B)	0	12	8,04	2,63	0,079	8,00	0,670	12
	schwierig (C)	0	12	5,76	2,55	0,077	6,00	0,480	12
	alle	9	36	24,55	5,68	0,171	25,00	0,682	36

N= 1100 je Bedingung.

C.2 *Part-whole-korrigierte* Trennschärfen der Punktauszahlungen

Tabelle C.2-1: *Part-whole-korrigierte* Trennschärfen der Items, berechnet über ausgezahlte Punkte für die Bedingungen „ME-linear 0 bis 100 Punkte-2 Antwortoptionen“ und „ME-linear 0 bis 100 Punkte-3 Antwortoptionen“.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-linear 0 bis 100 Punkte 2 Antwortoptionen	ME-linear 0 bis 100 Punkte 3 Antwortoptionen
1 bis 6	A1	3	0,16	0,14
1 bis 6	A1	4	0,25	0,29
1 bis 6	A1	9	0,25	0,31
1 bis 6	A1	13	0,42	0,44
1 bis 6	A1	15	0,38	0,44
1 bis 6	A1	19	0,19	0,23
7 bis 12	A2	35	0,32	0,32
7 bis 12	A2	36	0,43	0,46
7 bis 12	A2	37	0,32	0,29
7 bis 12	A2	38	0,37	0,39
7 bis 12	A2	39	0,15	0,29
7 bis 12	A2	56	0,38	0,41
13 bis 18	B1	67	0,38	0,42
13 bis 18	B1	72	0,54	0,57
13 bis 18	B1	73	0,21	0,30
13 bis 18	B1	75	0,24	0,35
13 bis 18	B1	76	0,34	0,38
13 bis 18	B1	79	0,27	0,40
19 bis 24	B2	82	0,35	0,44
19 bis 24	B2	85	0,33	0,34
19 bis 24	B2	89	0,12	0,38
19 bis 24	B2	90	0,44	0,56
19 bis 24	B2	91	0,15	0,35
19 bis 24	B2	95	0,21	0,46
25 bis 30	C1	98	0,47	0,43
25 bis 30	C1	100	0,33	0,37
25 bis 30	C1	101	0,34	0,32
25 bis 30	C1	110	0,25	0,36
25 bis 30	C1	111	0,42	0,40
25 bis 30	C1	113	0,45	0,42
31 bis 36	C2	125	0,14	0,24
31 bis 36	C2	134	0,21	0,33
31 bis 36	C2	143	0,22	0,29
31 bis 36	C2	144	0,33	0,45
31 bis 36	C2	153	0,39	0,30
31 bis 36	C2	157	0,29	0,36

N=1100 je Bedingung.

Tabelle C.2-2: *Part-whole-korrigierte* Trennschärfen der Items, berechnet über ausgezahlte Punkte für die Bedingungen „ME-logarithmisch -300 bis 100 Punkte-2 Antwortoptionen“ und „ME-linear -300 bis 100 Punkte-3 Antwortoptionen“.

Position im Test	GER-Stufe	Item Datenbank Nr.	ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen
1 bis 6	A1	3	0,15	0,18
1 bis 6	A1	4	0,27	0,17
1 bis 6	A1	9	0,23	0,24
1 bis 6	A1	13	0,33	0,35
1 bis 6	A1	15	0,25	0,34
1 bis 6	A1	19	0,21	0,19
7 bis 12	A2	35	0,23	0,31
7 bis 12	A2	36	0,34	0,40
7 bis 12	A2	37	0,29	0,28
7 bis 12	A2	38	0,29	0,31
7 bis 12	A2	39	0,19	0,19
7 bis 12	A2	56	0,24	0,33
13 bis 18	B1	67	0,29	0,34
13 bis 18	B1	72	0,41	0,40
13 bis 18	B1	73	0,16	0,22
13 bis 18	B1	75	0,28	0,29
13 bis 18	B1	76	0,26	0,36
13 bis 18	B1	79	0,20	0,32
19 bis 24	B2	82	0,32	0,38
19 bis 24	B2	85	0,29	0,29
19 bis 24	B2	89	0,12	0,21
19 bis 24	B2	90	0,35	0,36
19 bis 24	B2	91	0,17	0,27
19 bis 24	B2	95	0,19	0,33
25 bis 30	C1	98	0,44	0,40
25 bis 30	C1	100	0,26	0,29
25 bis 30	C1	101	0,20	0,32
25 bis 30	C1	110	0,20	0,28
25 bis 30	C1	111	0,34	0,28
25 bis 30	C1	113	0,41	0,39
31 bis 36	C2	125	0,10	0,20
31 bis 36	C2	134	0,21	0,25
31 bis 36	C2	143	0,25	0,29
31 bis 36	C2	144	0,29	0,35
31 bis 36	C2	153	0,36	0,30
31 bis 36	C2	157	0,32	0,40

N=1100 je Bedingung.

Tabelle C.2-3: *Part-whole-korrigierte* Trennschärfen der Items, berechnet über ausgezahlte Punkte für die Bedingungen mit Multiple-Choice „MC-0 oder 1 Punkt-2 Antwortoptionen“ und „MC-0 oder 1 Punkt-3 Antwortoptionen“.

Position im Test	GER-Stufe	Item Datenbank Nr.	MC-0 oder 1 Punkt 2 Antwortoptionen	MC-0 oder 1 Punkt 3 Antwortoptionen
1 bis 6	A1	3	0,18	0,19
1 bis 6	A1	4	0,20	0,24
1 bis 6	A1	9	0,25	0,34
1 bis 6	A1	13	0,35	0,38
1 bis 6	A1	15	0,33	0,39
1 bis 6	A1	19	0,15	0,18
7 bis 12	A2	35	0,29	0,32
7 bis 12	A2	36	0,38	0,40
7 bis 12	A2	37	0,23	0,30
7 bis 12	A2	38	0,32	0,32
7 bis 12	A2	39	0,17	0,26
7 bis 12	A2	56	0,26	0,31
13 bis 18	B1	67	0,32	0,40
13 bis 18	B1	72	0,51	0,54
13 bis 18	B1	73	0,13	0,24
13 bis 18	B1	75	0,20	0,29
13 bis 18	B1	76	0,28	0,37
13 bis 18	B1	79	0,23	0,35
19 bis 24	B2	82	0,30	0,47
19 bis 24	B2	85	0,26	0,27
19 bis 24	B2	89	0,13	0,27
19 bis 24	B2	90	0,34	0,44
19 bis 24	B2	91	0,17	0,26
19 bis 24	B2	95	0,15	0,39
25 bis 30	C1	98	0,43	0,35
25 bis 30	C1	100	0,16	0,30
25 bis 30	C1	101	0,31	0,28
25 bis 30	C1	110	0,29	0,34
25 bis 30	C1	111	0,32	0,27
25 bis 30	C1	113	0,37	0,33
31 bis 36	C2	125	0,09	0,16
31 bis 36	C2	134	0,24	0,28
31 bis 36	C2	143	0,20	0,15
31 bis 36	C2	144	0,32	0,39
31 bis 36	C2	153	0,26	0,19
31 bis 36	C2	157	0,26	0,27

N=1100 je Bedingung.

C.3 Deskriptive Statistik der Punktsummen nach einer Realismuskorrektur

Tabelle C.3-1: Realismuskorrigierte Punktsummen über alle Schwierigkeitsstufen.

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	1797	3600	2424,7	471,1	14,2	2349,5	67,4	73,8
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	1188	3600	2126,3	581,3	17,5	2054,0	59,1	68,0
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-10	3600	1341,3	929,1	28,0	1153,0	37,3	31,4
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-17	3600	1369,8	849,3	25,6	1274,5	38,1	22,4

N=1100 je Bedingung.

Tabelle C.3-2: Realismuskorrigierte Punktsummen der einfachen Items (GER-Stufe A).

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	570	1200	913,8	173,4	5,2	930,0	76,1	89,1
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	391	1200	870,8	202,3	6,1	901,0	72,6	89,5
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-105	1200	759,3	344,5	10,4	828,0	63,3	71,7
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-31	1200	821,6	315,7	9,5	912,0	68,5	71,7

N=1100 je Bedingung.

Tabelle C.3-3: Realismuskorrigierte Punktsummen der mittelschwierigen Items (GER-Stufe B).

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/-Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	474	1200	804,7	175,9	5,3	770,0	67,1	73,0
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	319	1200	711,0	233,0	7,0	683,5	59,3	67,0
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-301	1200	408,6	385,9	11,6	323,0	34,1	24,4
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-327	1200	418,2	378,1	11,4	364,5	34,9	11,4

N=1100 je Bedingung.

Tabelle C.3-4: Realismuskorrigierte Punktsummen der schwierigen Items (GER-Stufe C).

Bedingung	Minimum	Maximum	Mittelwert	Standardabweichung	Standardfehler	Median	Mittelwert/ Item	Mittelwert/ Item nicht korrigiert
ME-linear 0 bis 100 Punkte 2 Antwortoptionen	476	1200	706,2	163,0	4,9	639,0	58,9	59,3
ME-linear 0 bis 100 Punkte 3 Antwortoptionen	287	1200	544,4	194,5	5,9	474,5	45,4	47,6
ME-logarithmisch -300 bis 100 Punkte 2 Antwortoptionen	-502	1200	173,4	350,0	10,6	61,5	14,5	-2,0
ME-logarithmisch -300 bis 100 Punkte 3 Antwortoptionen	-432	1200	129,9	310,0	9,3	30,0	10,8	-16,1

N=1100 je Bedingung.

D. Auswahl der Distraktoren des dritten Experiments

Tabelle E-1: Auswahl des Distraktors für die Items mit zwei Antwortoptionen auf der Basis des Antwortverhaltens in den Multiple-Choice-Bedingungen der ersten beiden Experimente. N erstes Experiment=804, N zweites Experiment=808.

		Prozent Auswahl Distraktor 1			Prozent Auswahl Distraktor 2			Ausgewählter Distraktor
DB-Item Nr.	GER-Stufe	zweites Experiment	erstes Experiment	Mittelwert	zweites Experiment	erstes Experiment	Mittelwert	drittes Experiment
3	A1	0,4	0,2	**0,3**	3,6	1,5	**2,6**	2
4	A1	0,0	0,1	**0,1**	8,3	5,6	**7,0**	2
9	A1	4,0	5,3	**4,7**	4,8	3,9	**4,4**	1
13	A1	15,1	17,2	**16,2**	2,1	2,5	**2,3**	1
15	A1	3,3	2,9	**3,1**	7,9	8,3	**8,1**	2
19	A1	1,9	1,5	**1,7**	4,2	3,5	**3,9**	2
35	A2	12,3	10,1	**11,2**	8,7	7,3	**8,0**	1
36	A2	19,7	19,0	**19,4**	7,3	4,9	**6,1**	1
37	A2	3,2	3,5	**3,4**	11,9	11,9	**11,9**	2
38	A2	7,9	9,3	**8,6**	2,5	1,9	**2,2**	1
39	A2	14,1	13,4	**13,8**	5,1	5,2	**5,2**	1
56	A2	9,9	11,8	**10,9**	5,3	4,7	**5,0**	1
67	B1	21,4	21,8	**21,6**	12,0	12,9	**12,5**	1
72	B1	11,9	12,7	**12,3**	27,1	27,2	**27,2**	2
73	B1	13,7	15,0	**14,4**	4,3	4,5	**4,4**	1
75	B1	17,1	18,0	**17,6**	5,1	7,3	**6,2**	1
76	B1	23,9	21,1	**22,5**	11,0	12,4	**11,7**	1
79	B1	22,8	19,4	**21,1**	12,7	12,1	**12,4**	1
82	B2	17,6	19,8	**18,7**	10,5	9,2	**9,9**	1
85	B2	20,0	17,2	**18,6**	19,6	19,8	**19,7**	2
89	B2	15,6	14,9	**15,3**	23,5	25,6	**24,6**	2
90	B2	28,6	29,6	**29,1**	20,5	22,5	**21,5**	1
91	B2	20,0	18,8	**19,4**	34,3	34,5	**34,4**	2
95	B2	20,5	21,0	**20,8**	29,8	28,6	**29,2**	2
98	C1	13,0	12,1	**12,6**	28,8	33,2	**31,0**	2
100	C1	19,3	19,9	**19,6**	21,4	23,3	**22,4**	2
101	C1	15,1	15,2	**15,2**	36,5	37,9	**37,2**	2
110	C1	41,8	45,9	**43,9**	16,6	16,5	**16,6**	1
111	C1	26,1	24,9	**25,5**	21,2	25,2	**23,2**	1
113	C1	40,7	41,3	**41,0**	10,3	13,2	**11,8**	1
125	C2	41,7	38,0	**39,9**	13,2	15,5	**14,4**	1
134	C2	34,9	32,9	**33,9**	31,4	33,1	**32,3**	1
143	C2	30,6	30,0	**30,3**	35,3	36,8	**36,1**	2
144	C2	29,2	33,8	**31,5**	36,8	33,1	**35,0**	2
153	C2	34,7	34,1	**34,4**	26,9	22,2	**24,6**	1
157	C2	25,4	25,0	**25,2**	31,2	26,9	**29,1**	2

E. Texte der Übungen

Übungsitems unter Verwendung des Dreiecks im ersten Experiment:

1. Übung (von 5):
Stellen Sie sich vor, Sie hätten die Aufgabe, unter 3 englischen Marmeladenproben diejenige mit dem höchsten Ingwergehalt herauszuschmecken. Auf welches Feld des Dreiecks müssen Sie klicken, wenn Sie sich vollkommen (also zu 100%) sicher sind, dass "Probe B" den höchsten Ingwergehalt hat?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B
C: Marmeladenprobe C

2. Übung (von 5):
Welches Feld würden Sie anklicken, wenn Sie auf die selbe Frage antworten wollen, dass Sie "Probe C" sicher ausschließen können, da diese viel weniger nach Ingwer schmeckt als die anderen beiden Proben, Sie sich aber überhaupt nicht zwischen "Probe A" und "Probe B" entscheiden können, da beide für Sie gleich nach Ingwer schmecken?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B
C: Marmeladenprobe C

3. Übung (von 5):
Nehmen wir an, dass für Sie alle Marmeladenproben gleich schmecken. Sie wissen also überhaupt nicht, für welche Probe Sie sich entscheiden sollen. Auf welches Dreiecksfeld sollten Sie in diesem Fall klicken?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B
C: Marmeladenprobe C

4. Übung (von 5):

Welches Dreiecksfeld müssen Sie wählen, wenn Sie sich relativ sicher sind, dass "Probe B" am meisten Ingwer enthält, Sie aber nicht völlig sicher sind, dass nicht auch "Probe A" den höchsten Ingwergehalt haben könnte, Sie aber "Probe C" sicher ausschließen können?

Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*

A: Marmeladenprobe A

B: Marmeladenprobe B

C: Marmeladenprobe C

5. Übung (von 5):

Welches Teilstück des Dreiecks wählen Sie, wenn Sie zwar "Probe B" für sehr wahrscheinlich als die Marmelade mit dem höchsten Ingwergehalt erschmecken, Sie aber nicht 100%ig sicher sind, dass nicht auch "Probe A" oder "Probe C" den höchsten Ingwergehalt haben könnten?

Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*

A: Marmeladenprobe A

B: Marmeladenprobe B

C: Marmeladenprobe C

Übungen unter Verwendung der Schieberegler mit drei Antwortoptionen im ersten, zweiten und dritten Experiment:

1. Übung (von 5):
Stellen Sie sich vor, Sie hätten die Aufgabe, unter 3 englischen Marmeladenproben diejenige mit dem höchsten Ingwergehalt herauszuschmecken. Welche Slidereinstellungen wählen Sie, wenn Sie sich vollkommen (also zu 100%) sicher sind, dass "Probe B" den höchsten Ingwergehalt hat?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B
C: Marmeladenprobe C

2. Übung (von 5):
Welche Slidereinstellungen wählen Sie, wenn Sie auf die selbe Frage antworten wollen, dass Sie "Probe C" sicher ausschließen können, da diese viel weniger nach Ingwer schmeckt als die anderen beiden Proben, sich aber überhaupt nicht zwischen "Probe A" und "Probe B" entscheiden können, da beide für Sie gleich nach Ingwer schmecken?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B
C: Marmeladenprobe C

3. Übung (von 5):
Nehmen wir an, dass für Sie alle Marmeladenproben gleich schmecken. Sie wissen also überhaupt nicht, für welche Probe Sie sich entscheiden sollen. Welche Sliderwerte sollten Sie in diesem Fall einstellen?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B
C: Marmeladenprobe C

4. Übung (von 5):

Welche Sliderwerte sollten Sie einstellen, wenn Sie sich relativ sicher sind, dass "Probe B" am meisten Ingwer enthält, Sie aber nicht völlig sicher sind, dass nicht auch "Probe A" den höchsten Ingwergehalt haben könnte, Sie "Probe C" aber sicher ausschließen können?

Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*

A: Marmeladenprobe A

B: Marmeladenprobe B

C: Marmeladenprobe C

5. Übung (von 5):

Welche Slidereinstellung wählen Sie, wenn Sie zwar "Probe B" für sehr wahrscheinlich als die Marmelade mit dem höchsten Ingwergehalt erschmecken, Sie aber nicht 100%ig sicher sind, dass nicht auch "Probe A" oder "Probe C" den höchsten Ingwergehalt haben könnten?

Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*

A: Marmeladenprobe A

B: Marmeladenprobe B

C: Marmeladenprobe C

Übungen unter Verwendung der Schieberegler mit zwei Antwortoptionen im dritten Experiment:

1. Übung (von 3):
Stellen Sie sich vor, Sie hätten die Aufgabe, unter 2 englischen Marmeladenproben diejenige mit dem höchsten Ingwergehalt herauszuschmecken. Welche Slidereinstellungen wählen Sie, wenn Sie sich vollkommen (also zu 100%) sicher sind, dass "Probe B" den höchsten Ingwergehalt hat?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt“?*
A: Marmeladenprobe A
B: Marmeladenprobe B

2. Übung (von 3):
Nehmen wir an, dass für Sie beide Marmeladenproben gleich schmecken. Sie wissen also überhaupt nicht, für welche Probe Sie sich entscheiden sollen. Welche Sliderwerte sollten Sie in diesem Fall einstellen?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B

3. Übung (von 3):
Wie stellen Sie die Slider ein, wenn Sie zwar "Probe B" relativ wahrscheinlich als die Marmelade mit dem höchsten Ingwergehalt erschmecken, Sie aber nicht völlig sicher sind, dass nicht auch "Probe A" den höchsten Ingwergehalt haben könnte?
Frage: *„Welche Marmeladenprobe hat den höchsten Ingwergehalt?“*
A: Marmeladenprobe A
B: Marmeladenprobe B

F. Anzahl der möglichen Verteilungen bei Schiebereglern

Berechnung der Anzahl der möglichen Verteilungen von Antwortsicherheit auf drei Antwortoptionen bei einer Erhebung von Antwortsicherheit mithilfe von Schiebereglern:

Definition der Menge P

P wird definiert als die Menge aller möglichen prozentualen Antwortsicherheiten je Antwortoption:

(1) $P := \{0,1,2,...,100\}$

Die Mächtigkeit (Anzahl der Elemente) der Menge P beträgt somit:

(2) $|P| = 101$

Definition der Menge S

S wird definiert als die Menge aller Tripel (x,y,z), bei denen x+y+z=100 ergibt:

(3) $S := \{(x, y, z) \in P^3 \mid x + y + z = 100\}$

Jedes Element dieser Menge entspricht einer möglichen Verteilung von Antwortsicherheit auf die drei Antwortoptionen. $|S|$, also die Mächtigkeit von S, entspricht damit der Gesamtzahl aller möglichen Verteilungen.

Definition von Teilmengen S_i

Die Menge S wird in 101 paarweise disjunkte (elementfremde) Teilmengen $S_i \subset S$ zerlegt, deren erste Komponente x jeweils einen festen Wert i besitzt:

(4) $S_i := \{(x, y, z) \in S \mid x = i\} \quad (i \in P)$

Die Vereinigungsmenge aller Teilmengen S_i ergibt wieder die Menge S:

(5) $\bigcup_i S_i = S$

Die Mächtigkeit $|S_i|$ einer jeden Teilmenge S_i beträgt:

(6) $|S_i|$=101-i

Begründung: Die Teilmengen enthalten jeweils folgende Tripel (x,y,z):
x=i (i ist fest)
y=0,1,2,…,100-i (y ist abhängig von x)
z=100-i-y (z ist abhängig von y)
Für y und z ergeben sich somit 100-i+1, also 101-i mögliche Werte.
Beispiel für i=95: S_{95}={(95,0,5), (95,1,4), (95,2,3), (95,3,2), (95,4,1), (95,5,0)}

Regeln und Formeln

Für paarweise disjunkte Mengen S_i gilt die Regel:

(7) $$\left|\bigcup_i S_i\right| = \sum_i |S_i|$$

Die Summe der ersten n aufeinander folgenden natürlichen Zahlen berechnet sich nach der Gaußschen Summenformel:

(8) $$\sum_{i=1}^{n} i = \frac{n}{2} * (n+1)$$

Berechnung der Mächtigkeit von S

$$|S|$$
$$= \left|\bigcup_i S_i\right| \qquad \text{laut (5)}$$
$$= \sum_{i=0}^{100} |S_i| \qquad \text{laut (7)}$$
$$= \sum_{i=0}^{100} (101 - i) \qquad \text{laut (6)}$$
$$= \sum_{i=0}^{100} 101 - \sum_{i=0}^{100} i$$
$$= 101 * 101 - \frac{100}{2} * (100+1) \qquad \text{laut (8)}$$
$$= \underline{5151}$$

Es gibt also 5151 unterschiedliche Verteilungen von Antwortsicherheit auf drei Antwortoptionen.

Erklärung zur Dissertation

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 12.09.2011

(Erika Heidi Enders)