U N I V E R S I T Ä T D U I S B U R G - E S S E N
Fachbereich Germanistik

TESTKLUGHEIT TESTEN

Die Tücken der Multiple-Choice-Fragen

Praktische Hausarbeit

Aufgabenentwicklung im Deutschunterricht

Dr. Ulrike Behrens

Wintersemester 2009/2010

Von: Franziska Rieke

Matrikelnummer: 2236348

4. Fachsemester

Inhaltsverzeichnis

1. Einleitung

Der Mensch gerät im Laufe seines Lebens immer wieder in Situationen, in denen er sein Wissen und Können einer Leistungskontrolle unterziehen muss. Durch unzählige Klausuren in der Schulzeit wird er auf Testsituationen vorbereitet und so scheint es nicht verwunderlich, dass das Testen heutzutage auch zum Freizeitspaß werden kann. Schnell werden in der Mittagspause ein paar Kreuze in einem Frauenmagazin gesetzt, um herauszufinden zu welchem „Typ Frau“ man gehört.

Doch was verbirgt sich hinter diesen Multiple-Choice-Tests? Eine Frage mit vier möglichen Antwortmöglichkeiten zu Papier zu bringen scheint auf den ersten Blick einfach und nicht sehr aufwändig. Doch oft entpuppt sich diese Aufgabe als ein kleines Kunstwerk, da viele Fallen auf dem Weg zu einer guten Multiple-Choice-Frage lauern. In dieser Arbeit widme ich mich dem Testverfahren durch Multiple-Choice-Tests, um aufzuzeigen, welche Probleme bereits bei der Erstellung solcher Tests entstehen können und inwieweit diese später die Ergebnisse der Testpersonen beeinflussen.

Im Anschluss an die Erörterung der Tücken einer Multiple-Choice-Frage wird eine Umfrage mit Hilfe eines kurzen Fragebogens durchgeführt. Dieser enthält zehn Multiple-Choice-Fragen, die lediglich dazu entwickelt wurden die „Testklugheit“ der Testteilnehmer zu testen. Der Fragebogen kann Lehrern und anderen Personen, die selbst einen Multiple-Choice-Fragebogen erstellen müssen, als Negativ-Vorbild dienen, indem er ihnen schnell die möglichen Fehler aufzeigt, die dabei begangen werden können. Aber auch die Testpersonen lernen anhand dieses Fragebogens schnell, dass es Strategien gibt ohne jegliches Wissen zu der richtigen Antwortoption zu gelangen.

2. Multiple Choice Tests (MC-Tests)

2.1 Aufbau von MC-Fragen

MC-Aufgaben sind auf den ersten Blick recht schlicht aufgebaut. Auf eine Frage (Itemstamm) folgen zwischen zwei und fünf Antwortoptionen (Alternativen). Von diesen Optionen ist meist nur eine korrekt. Die falschen Lösungen werden als Distraktoren bezeichnet. Der Testperson ist pro MC-Frage in der Regel nur das Ankreuzen einer Lösungsmöglichkeit gestattet. Sollte sie mehrere Optionen ankreuzen, so muss die Lösung als falsch gewertet werden.

2.1.1 Itemstamm

Der Itemstamm kann, wie in folgendem Beispiel, eine geschlossene Frage bilden, auf die vier mögliche Antwortoptionen vorgegeben werden.

Welcher Vorgang findet in der Landwirtschaft statt?

- ☐ Das Feld wird bestellt.
- ☐ Das Feld wird versandt.
- ☐ Das Feld wird angenommen.
- ☐ Das Feld wird ausgepackt.

Die Frage kann allerdings auch aus einer unvollständigen Aussage bestehen, die weitergeführt werden muss.

In der Landwirtschaft wird ein Feld...

- ☐ bestellt.
- ☐ versandt.
- ☐ angenommen.
- ☐ ausgepackt.

Nach Gronlund (2006, 75) ist das Aufgabenformat als unvollständiger Satz prägnanter, da es weniger Text umfasst. Jedoch ist es für den Aufgabenentwickler einfacher die Aufgabenstämme als Fragen zu konstruieren. Daher empfiehlt sich für ungeübte Aufgabenent-

wickler, die Form der geschlossenen Frage zu wählen und erst später mit zu vervollständigenden Sätzen zu arbeiten.

Die Fragestellung sollte stets positiv formuliert werden, da Verneinungen schnell überlesen werden können und dadurch nicht mehr das Wissen, sondern die Leseaufmerksamkeit der Testperson überprüft wird. Gibt es keine andere Möglichkeit als die Verneinung einer Frage, so muss die Verneinung durch Fettdruck und/oder Unterstreichungen deutlich gekennzeichnet werden (vgl. Bremerich-Vos/Köller 2007, 21). Auch sollte der Itemstamm immer möglichst klar und in verständlicher Sprache formuliert werden. Ist er nur schwer verständlich, kann die erbrachte Leistung der Testperson nicht mit ihrem vorhandenen Wissen gleichgesetzt werden, da zumindest partiell Lesekompetenz (hier: Verstehen der Fragestellung) im Spiel ist. Dies widerspricht dem Ziel, genau die angezielten Fähigkeiten/Kenntnisse (und keine anderen) zu messen.

Der Itemstamm hat des Weiteren oft die Aufgabe der präzisen und leicht verständlichen Problembeschreibung (vgl. Schmidts/Lischka 2001, 7). Beispielsweise wird bei medizinischen Tests häufig eine Fallbeschreibung vor die eigentliche Frage gestellt, die die Krankheitsverläufe oder aufgetretenen Symptome der Patienten beschreiben. Die Fallbeschreibung hilft der Testperson, die im Anschluss gestellte MC-Frage korrekt zu beantworten.

2.1.2 Antwortoptionen

Der wohl schwierigste Vorgang bei der Erstellung einer MC-Aufgabe ist die Formulierung der Antwortalternativen. Meist werden pro Frage jeweils vier (seltener fünf oder drei) Alternativen formuliert, von denen meist nur eine richtig ist. Die gesetzten Distraktoren (also die falschen Antwortoptionen) müssen bewusst entwickelt und mit Bedacht ausgewählt werden, denn sie entscheiden sowohl über die Schwierigkeit als auch über die Verständlichkeit der Aufgabe. Damit die richtige Lösung unter den Distraktoren nicht auffällt, müssen sie möglichst plausibel klingen. Sie dürfen die Testteilnehmer, die die richtige Lösung kennen, nicht verunsichern und müssen für sie eindeutig als falsch zu erkennen sein (vgl. Bremerich-Vos/Köller 2007, 21). Sie müssen grammatikalisch an die Frage angepasst

sein, dürfen in Länge und Form nicht voneinander abweichen, keine Wörter enthalten, die auf eine falsche Antwort schließen lassen (immer, ausschließlich, nie, etc.) und sich nicht in direktem Wortlaut auf die Fragestellung beziehen. Um geeignete Distraktoren zu finden empfiehlt es sich, die MC-Fragen im Rahmen der Testentwicklung als offene Fragen zu formulieren und von einer Testgruppe beantworten zu lassen. Aus den frei formulierten Antworten, können daraufhin falsche Antworten als Distraktoren ausgewählt werden. Auf diese Weise ist die Plausibilität der Distraktoren aus Sicht der Testpersonen sichergestellt (vgl. Jacobs 2005).

Auch die Reihenfolge der richtigen Antworten bei mehreren aufeinander folgenden MC-Items muss sinnvoll ausgewählt werden. Ergeben die Lösungen eindeutige Muster oder folgt die Lösungsfolge einem System, welches von der Testperson durchschaut werden kann, kann dies zu verfälschten Leistungsergebnissen führen. Beispielsweise können Testpersonen, die die richtige Antwort nicht kennen, sich über ein auffälliges Ankreuzschema die korrekte Antwort erschließen. Hingegen sind Testpersonen, die die korrekte Antwort kennen, möglicherweise bei einer Abweichung vom vorigen Schema irritiert und neigen dazu, die falsche Antwort zu geben. Daher sollte die Reihenfolge der richtigen Antwortmöglichkeit stets nach dem Zufallsprinzip gewählt werden (vgl. Grolund 2006, 89).

Zwischen den Distraktoren und der richtigen Lösung gibt es zwei Abgrenzungsverhältnisse. Entweder ist die korrekte Lösung wahr und die Distraktoren sind falsch („true answer form“) oder die korrekte Lösung ist die beste der vier Antwortalternativen („best answer form“) (vgl. Grolund 2006, 82). Auch bei der „best answer form“ muss allerdings eine Lösung existieren, die von der Testperson mit Abstand am logischsten und sinnvollsten erkannt werden kann und die somit als richtige Antwort zu identifizieren ist. Welche dieser Formen im Test angewandt wird, sollte in einem kurzen Zusatz im Itemstamm erläutert werden.

2.2 Testgütekriterien

Damit der zu erstellende Test den Testgütekriterien entspricht, muss er objektiv, reliabel, und valide sein (vgl. Lienert/ Raatz 1994, 6).

- **Objektivität**

Objektiv ist ein Test dann, wenn Testdurchführung und Testergebnisse der Teilnehmer voneinander unabhängig sind (vgl. Lienert/Raatz 1994, 7). Die Ergebnisse müssen demnach unabhängig vom Untersucher stets identisch sein. Hier lässt sich zwischen drei Aspekten unterscheiden:

‣ Durchführungsobjektivität

Dieser Aspekt betrifft den Zusammenhang zwischen den Testergebnissen und den (Verhaltens-)Varianten des Testleiters bzw. der Testsituation. Unterschiede im Verhalten des Testleiters beeinflussen die Testteilnehmer und verändern ihre Ergebnisse. Um eine sehr hohe Durchführungsobjektivität zu erhalten, empfiehlt es sich, die Interaktion zwischen dem Testteilnehmer und dem Testleiter auf ein Minimum zu reduzieren. Daher sollten Durchführungsinstruktionen so genau wie möglich in schriftlicher Form vorgelegt und die Situation der Untersuchung standardisiert werden (vgl. Lienert/Raatz 1994, 8). Bei dem für diese Arbeit entwickelten Fragebogen, der per E-Mail verschickt wurde, fand kein persönlicher Kontakt statt; die Durchführungsinstruktionen waren für alle Testpersonen identisch. Da die Durchführung jedoch nicht durch eine Testleitung kontrolliert wurde, waren die Rahmenbedingungen der Testdurchführung sehr unterschiedlich. Die Testpersonen lösten den Fragebogen teilweise in Gruppen, wodurch beispielsweise die aufgewandte Zeit zur Lösung der Fragen stark variierte.

‣ Auswertungsobjektivität

Die Auswertung der Fragebögen muss nach vorgegeben Regeln erfolgen, damit die Leistung gerecht bewertet werden kann (vgl. Lienert/Raatz 1994, 8). Bei dem hier entwickelten Multiple-Choice-Test ist die Auswertungsobjektivität dadurch gegeben, dass es weder in der Beantwortung noch in der Bewertung Spielräume gibt. Die „richtige“ Antwortoption ergibt sch aus der theoriegeleiteten Manipulierung (s.u.). Nur wer diese ankreuzt, verhält sich im Sinne der Hypothese. Eine niedrigere Auswertungsobjektivität wäre bei freien Aufgaben-

formaten gegeben, bei denen es viel schwieriger ist, die Lösungen unter vorgegebenen Regeln zu beurteilen und somit objektiv zu bewerten.

‣Interpretationsobjektivität

Nachdem der Test ausgewertet wurde, ist es wichtig, dass das Ergebnis von verschiedenen Testteilnehmern gleich gedeutet werden kann. In dem für diese Arbeit entwickelten Fragebogen ist dies gegeben: Zwar werden hier keine Kenntnisse überprüft, die erreichte bestimmte Gesamtpunktzahl kann aber als Maß der „Testklugheit“ interpretiert werden.

• **Reliabilität**

Unter der Reliabilität eines Tests versteht man die Genauigkeit bzw. Verlässlichkeit, mit der er ein Leistungsmerkmal bestimmt (vgl. Lienert/Raatz 1994, 9). Dabei bilden sie den Teil der Ergebnisse, der nicht von Messfehlern beeinträchtigt wird. Reliable Testverfahren sind absolut zuverlässig und stets zu wiederholen, ohne eine Abweichung des Testergebnisses zu erhalten. Dabei gilt, dass nur ein Test, der objektiv ist, auch reliabel sein kann (vgl. Bremerich-Vos/Köller 2007, 4).

• **Validität**

Unter der Validität versteht man die Gültigkeit eines Testes. Die höchste Validität ist erreicht, wenn das Ergebnis eines Tests einen fehlerfreien Rückschluss auf die Leistungsfähigkeit der Testperson ermöglicht (vgl. Lienert/Raatz 1994, 10). Auch hier zeigt die Intensität der Auseinandersetzung der Testpersonen mit den entwickelten MC-Fragen Effekte, da durch den mangelnden Anreiz die Strukturen zu durchschauen nicht zwangsläufig ein schlechtes Testergebnis auf mangelnde Leistungsfähigkeit schließen lässt, sondern vielleicht eher auf wenig Engagement, ungünstige Rahmenbedingungen o.ä. Sollte sich dies bestätigen, so wäre der Test nicht valide.

3. Testentwicklung und Testdurchführung

Der hier vorgestellte Multiple-Choice-Test entstand nach einer amerikanischen Vorlage („Test your Testwiseness“[1]). Ziel ist es, Auswirkungen der Itemkonstruktion auf das Ankreuzverhalten der Testpersonen nachzuweisen. Da diese vollkommen unabhängig von den Kenntnissen der Probanden sein sollen, wird mit Kunstwörtern operiert. Es ist also unmöglich, die „richtige“ Lösung auf Basis von Vorwissen zu finden.

3.1 Entwicklung des Fragebogens

Die richtige Lösung einer MC-Frage zu finden ist nicht immer leicht. Doch häufig werden Fehler beim Erstellen der Fragen begangen, die der Testperson helfen, auch ohne das nötige Wissen die richtige Antwort zu finden. Um den Einfluss der Formulierung und Struktur von MC-Fragen auf die Wahl der Antwortoptionen zu untersuchen, habe ich einen Fragebogen entworfen, der darauf abzielt, allein durch den Itemstamm und die Antwortoptionen das Verhalten der Testpersonen bei der Beantwortung der Fragen zu beeinflussen. Das Fach- und Weltwissen der Testpersonen darf bei der Beantwortung keinen Einfluss auf ihre Antwort haben, da dieses die Testpersonen dazu verleiten würde, lediglich fachlich korrekte Antworten als richtige Lösung anzusehen.

Die Wörter der deutschen Sprache sind für jeden Menschen mit individuellen Assoziationen konnotiert. Dadurch ist es sehr wahrscheinlich, dass die Antworten der Testperson auch von diesen Assoziationen beeinflusst werden. Da diese persönlichen Konnotationen in der Testauswertung nicht nachvollzogen werden können, wird im Fragebogen mit erfunden Wörtern operiert, die keinerlei Verbindung mit bekannten deutschen Wörtern und dazugehörigen Konnotationen zulassen. Dadurch besteht andererseits die Gefahr, dass die Testpersonen sich aufgrund der unbekannten Wörter schnell überfordert fühlen und nicht mehr versuchen, die Strukturen, die hinter den Fragen stehen, zu durchschauen. Jedoch liegt jeder Frage eine Struktur zugrunde, die die Testpersonen dazu verleiten soll, eine bestimmte Option als einzig logische Antwort anzusehen. Diese provozierten Antworten ba-

[1] online verfügbar unter http://www.emsc.nysed.gov/osa/assesspubs/pubsarch/ActivityTestYourTestwiseness.pdf

sieren teilweise auf der Verwendung der deutschen Grammatik, welcher jeder Testteilnehmer mächtig sein muss, um erwartungsgemäß zu antworten.

Die Verwendung einer Phantasiewörter dient darüber hinaus dazu, inhaltliche Aspekte der Fragen, die das Wissen der Testpersonen testen, ausklammern zu können. In angewandten Testfragen, beispielsweise in der Schule, sollte dennoch jede Frage stets ein wichtiges Lernziel abfragen (vgl. Grolund 2006, 77) und die Frage nach dem Inhalt und Ziel der zu entwickelnden MC-Frage eine der ersten Fragen sein, die sich der Testentwickler stellt.

Insgesamt erwies sich die Entwicklung des in erfundener Sprache verfassten Fragebogens als eine Aufgabe voller Tücken. Denn oft verwendete ich intuitiv gleich klingende oder gar identische Wörter, die ungewollt Einfluss auf die Beantwortung der Testpersonen hätten nehmen können. Daher erstellte ich den Fragebogen zuerst in deutscher Sprache und änderte anschließend die Fragen so ab, dass sie unverständlich wurden.

In den folgenden sechs Teilkapiteln werden die möglichen und verbreiteten Fehler erläutert, welche in dem entwickelten Fragebogen erforscht wurden.

3.1.1 Wiederholung von Satzteilen aus der Fragestellung

Schnell ist eine Frage so gestellt, dass sie Wörter aus der richtigen Antwort enthält. Wird die Testpersonen dadurch dazu verleitet, die Antwort mit den kongruierenden Wörtern anzukreuzen? Sollte dies der Fall sein, werden MC-Fragen daher im besten Fall so formuliert, dass in keiner Weise Wörter der Frage in den Antwortoptionen benutzt werden. In dem erstellten Fragebogen wird diese Frage durch zwei Testaufgaben, deren Antworten jeweils ein Wort aus der Frage beinhalten, getestet.

Der Flisker in Graparten entsteht durch...

- ☐ Gloksen-Iskiner
- ☐ Hipanen-Jolader
- ☐ Flisker-Haptakten
- ☐ Wosten-Redunen

Lösung: Flisker-Haptakten

Durch die Wiederholung des Nomen „Flisker“ ist die einzige sinnvolle Antwort „Flisker-Haptakten“. Die Antwort zitiert direkt das Subjekt der Frage.

Herpetani finden statt, wenn...

- ☐ ...Gromanen herpetanen.
- ☐ ...Lopsten raden und die Flurale stimmt.
- ☐ ...Uganuren sich grabsten.
- ☐ ...Schnobel flochsieren.

Lösung: ... Gromanen herpetanen

Bei dieser Aufgabe wird das Subjekt abgewandelt und in der richtigen Antwort als Prädikat verwendet.

3.1.2 Qualifizierende Beiwörter

In MC-Fragen werden oft Wörter wie „manchmal“, „normalerweise“ oder „teilweise“ verwendet. Führt dies die Testpersonen dazu, diese Antworten als richtig und im Gegenzug absolute Formulierungen wie „immer“ oder „nie“ als einen Hinweis auf falsche Antworten anzusehen? (Vgl. Bremerich-Vos/Köller 2007, 22) Zur Überprüfung dieser Annahme wird folgende Frage im MC-Test gestellt:

Ein Gor gewinnt gegen den Hemni, da...

- ☐ ...kein Hemni stolk ist.
- ☐ ...Goren meistens sehr flop sind.
- ☐ ...Goren immer stimpen.
- ☐ ...alle Hemni dompeln

Lösung: ... Goren meistens sehr flop sind.

Wörter wie „kein“, „immer“ und „alle“ sind Hinweise auf Distraktoren. Hingegen soll ein Wort wie „meistens“ auf die richtige Antwort hinweisen.

3.1.3 Gekoppelte Fragen

Bei diesem Phänomen sind zwei Fragen insofern voneinander abhängig, als die eine Frage Hinweise auf die Antwort der anderen Frage gibt. Schafft die Testperson es, die einzelnen Aufgaben in Verbindung zu bringen, so kann die Lösung der einen Aufgabe der Aufgabenstellung einer weiteren Aufgabe entnommen werden. Daher sollten die Fragen von

dem Testentwickler nicht nur einzeln, sondern auch im Zusammenhang kritisch betrachtet werden.

Die wichtigste Funktion eines Pflopfes kann nur in Verbindung mit welchem Gegenstand ausgeübt werden?

- ☐ Mit der moleren Quaterne
- ☐ Mit einem exinen Pavanus
- ☐ Mit dem Ropf eines Fertuns
- ☐ Mit einem hesterinaten Groller

Lösung 1: ... mit einem hesterinaten Groller

E Womit wird der hesterinate Groller geöffnet?

- ☐ Einem Pflopf
- ☐ Einer Kierke
- ☐ Einer Lierse
- ☐ Einem Redel

Lösung 2: Einem Pflopf

Die Erwähnung des „Pflopfes“ und des „hesterinaten Grollers“ in den jeweiligen Fragestellungen sollen der Testperson Hinweise auf die richtige Antwort geben.

3.1.4 Grammatikalische Strukturen

Achtet eine Testperson bei der Beantwortung einer MC-Frage auf die Grammatik? Wenn dies der Fall sein sollte, kann sie in dieser Aufgabe drei der vier Optionen ausschließen, da sie grammatikalisch nicht zur Fragestellung passen. Insbesondere bei der Formulierung der Frage als offener Satz, der weitergeführt werden muss, kann sich dieser Fehler schnell einschleichen.

Die Ursachen für die Falkverwangung sind...

- ☐ ...die Zottler und das Gasperatum.
- ☐ ...ein Holk im Geranium.
- ☐ ...Hestina mit Verstinden spielen.
- ☐ ...einige Walkos haben Trasper verloren.

Lösung: ... die Zottler und das Gasperatum

Durch die grammatikalisch falschen Strukturen der Distraktoren (Singular, Satzbau) können diese ausgeschlossen werden.

3.1.5 Abstimmung der Antwortoptionen

Nicht nur durch die Fragestellung können Distraktoren ausgeschlossen werden, sondern auch durch die Formulierung der Antwortoptionen. Dieser Formulierungsfehler muss nicht zwangsläufig zur richtigen Lösung führen, kann aber oft die Antwortoptionen um ein oder zwei Distraktoren reduzieren. Daher müssen alle Antwortoptionen so aufeinander abgestimmt werden, dass sie keine Auskunft über die richtige Lösung geben. Beschäftigt sich die Testperson intensiv mit den Antwortoptionen, so kann sie diese Frage leicht richtig beantworten.

Welche Testona gibt es **nicht**?

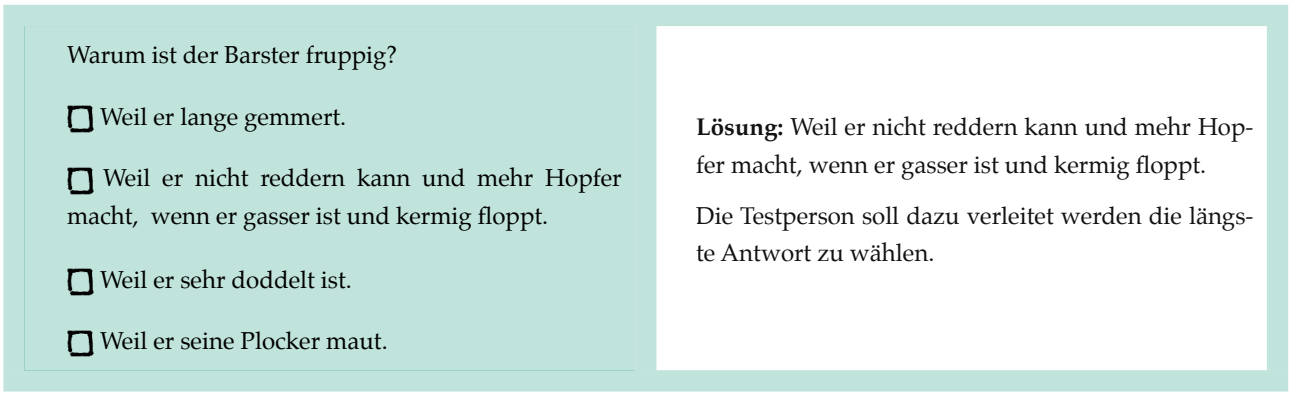
- ☐ Garst und Leonta
- ☐ Flocker und Leonta
- ☐ Leonta
- ☐ Plosper und Leonta

Lösung: Leonta

Die Antwort „Leonta“ ist in jeder Antwortoption inbegriffen, daher kann die Testperson folgern, dass „Leonta“ richtig sein muss. Über die anderen Begriffe wird keine Auskunft gegeben.

3.1.6 Länge der Antwortoptionen

Welchen Einfluss hat die Länge der Antwortoptionen? Werden die längeren Antworten den kürzeren gegenüber präferiert? Mit der hier entwickelten Aufgabe kann diese Frage getestet werden. Bestätigt sie sich, sollten die Antwortoptionen stets in der gleichen Länge und gleichen Ausführlichkeit formuliert werden.



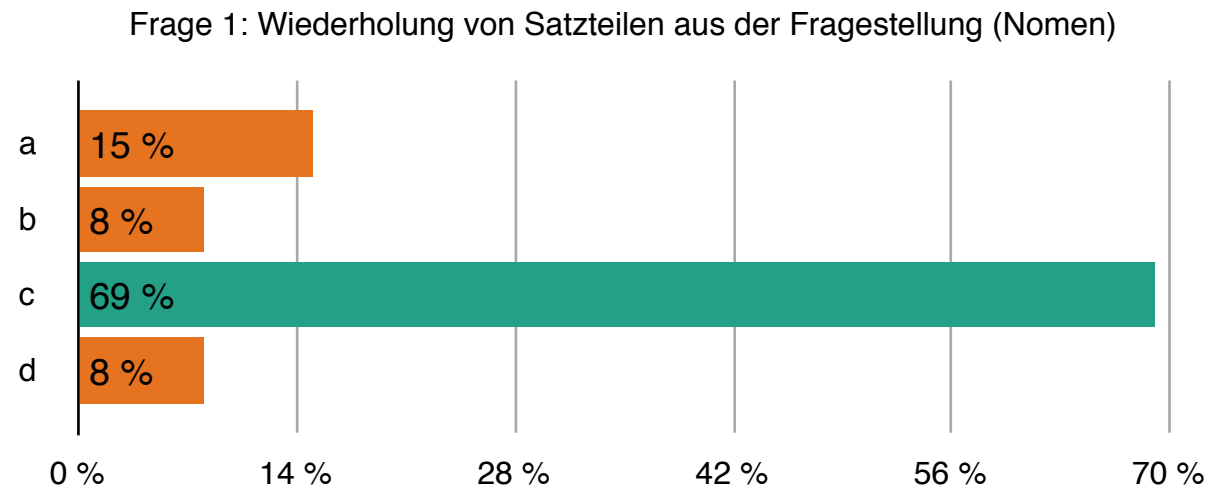
3.2 Durchführung

Um den Test mit möglichst vielen Teilnehmern durchführen zu können, verschickte ich eine E-Mail mit den Testinstruktionen und dem angehängten Test an circa 150 Freunde und Bekannte. Diese Gruppe setzte sich, neben einigen Schülern (unter 20 Jahren alt) und wenigen Berufstätigen (über 30 Jahre alt), hauptsächlich aus Studenten im Alter von 20 bis 30 Jahren zusammen, welche während ihres Studiums bereits Erfahrungen mit dem Multiple-Choice-Testverfahren sammeln konnten. Alle Testpersonen bekamen identische Testinstruktionen, welche durch die Mail konkret dargelegt wurden. Innerhalb weniger Tage bekam ich 52 ausgefüllten Fragebögen zurück. Einige Testteilnehmer schrieben in der Antwortmail noch ein kurzes Feedback zum Fragebogen. Aus diesem ging hervor, dass einige Probanden weniger Zeit und Aufwand in die Beantwortung der Fragen legten als andere. Manche kreuzten die Antworten an, die für sie am besten klangen und andere versuchten, unbedingt unter den Antwortoptionen eine Antwort als richtig zu entlarven. Manche Testpersonen lösten den Fragebogen zusammen mit ihren Freunden und Verwandten und nutzen ihn für ein unterhaltsames Abendprogramm. Dabei könnte der Austausch über die Testfragen in einem normalen Test zu Verfälschungen führen. Jedoch wer-

den in dem entwickelten Fragebogen keine Wissensfragen gestellt, in der nur eine Antwortoption als richtig gilt und vom Teilnehmer angekreuzt werden muss. Vielmehr werden Fragen gestellt, die bei der Testperson keine richtige Lösung, sondern eine logische Folgerung der Antwort provozieren sollen. Hierbei hat auch das gemeinsame Ausfüllen keine Auswirkung auf das Ergebnis.

4. Ergebnisse und Diskussion

4.1 Frage 1: Wiederholung von Satzteilen aus der Fragestellung (Nomen)



Der Flisker in Graparten entsteht durch...

☐ ...Gloksen-Iskiner (8)

☐ ...Hipanen-Jolader (4)

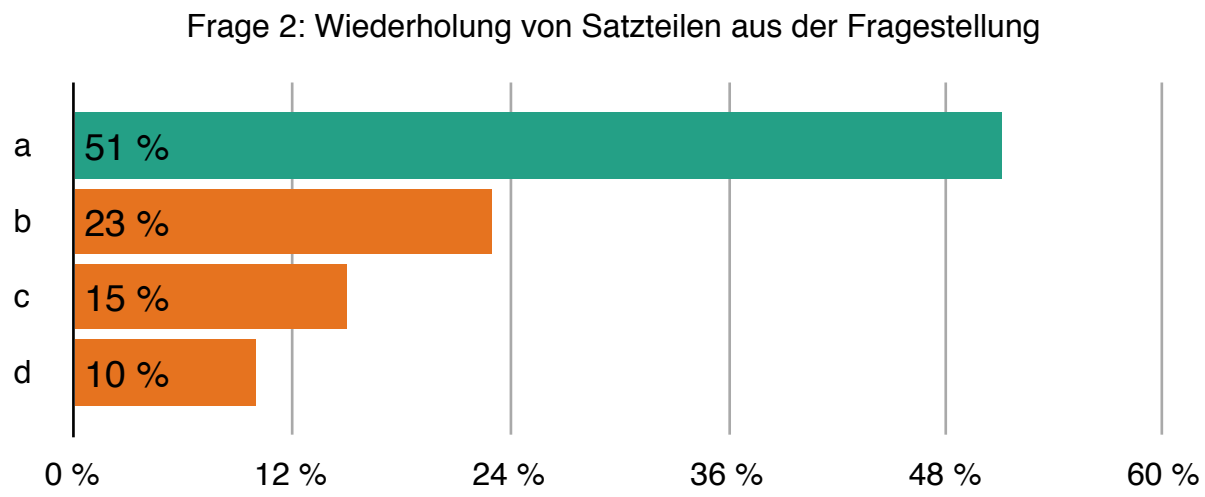
☑ **...Flisker-Haptakten (36)**

☐

Auswertung:

Das Ergebnis der ersten Frage fällt recht eindeutig aus. Die Testpersonen haben zu 69% die Antwort gewählt, die das Wort „Flisker“ aus der Fragestellung wieder aufnimmt. Es kann vermutet werden, dass der Großteil der Testpersonen die Struktur dieser Frage durchschaut oder diese Antwort zumindest unterbewusst ausgewählt hat. Somit ist die These, dass Antworten als korrekt angesehen werden, wenn sie Wörter aus der Fragestellung enthalten, bestätigt.

4.2 Frage 2: Wiederholung von Satzteilen aus der Fragestellung (Verb)



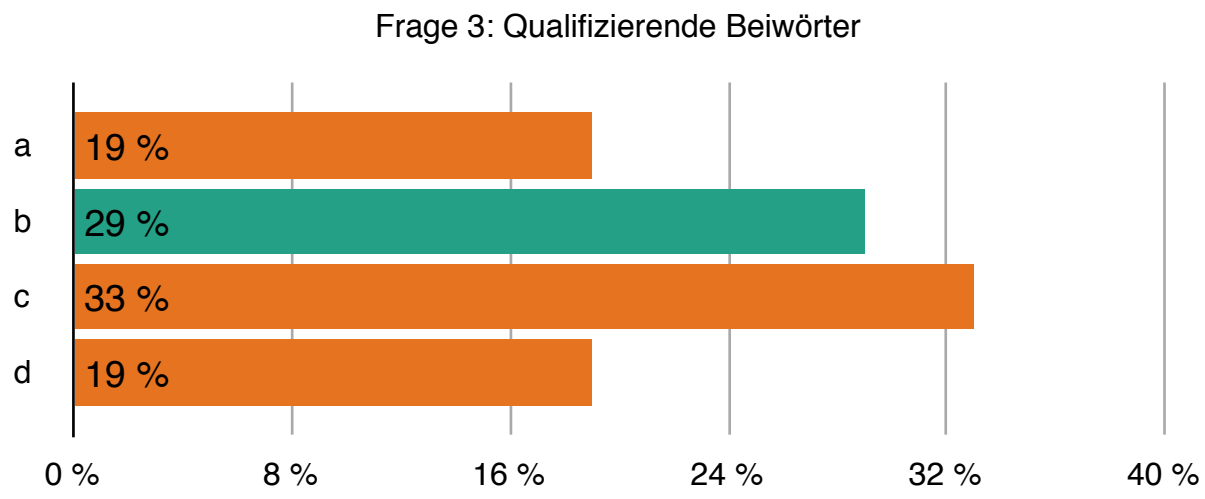
Herpetani finden statt, wenn...

- [x] **...Gromanen herpetanen (27)**
- [ ] ...Lopsten raden und die Flurale stimmt (12)
- [ ] ...Uganuren sich grabsten (8)
- [ ] ...Schnobel flochsieren (5)

Auswertung:

Im Unterschied zur ersten Frage liegt die Quote der durch die Itemkonstruktion provozierten Antworten bei der abgewandelten Form der ersten Frage nur knapp über 50%. Dies ergibt einen „Erfolgs"unterschied zwischen diesen beiden Fragen von 18%, obwohl sie nach dem gleichen Muster aufgebaut sind. In dieser Aufgabe wurde lediglich das Nomen „Herpetani" in das Verb „herpetanen" abgewandelt. Diese Verbindung fiel weniger Testpersonen auf als die direkte Wiederholung eines Wortes aus der Fragestellung. Dennoch ist die Quote der richtigen Antworten überzufällig häufig, denn die Lösungswahrscheinlichkeit liegt aufgrund der vier möglichen Optionen bei nur 25%.

4.3 Frage 3: Qualifizierende Beiwörter



Ein Gor gewinnt gegen den Hemni, da...

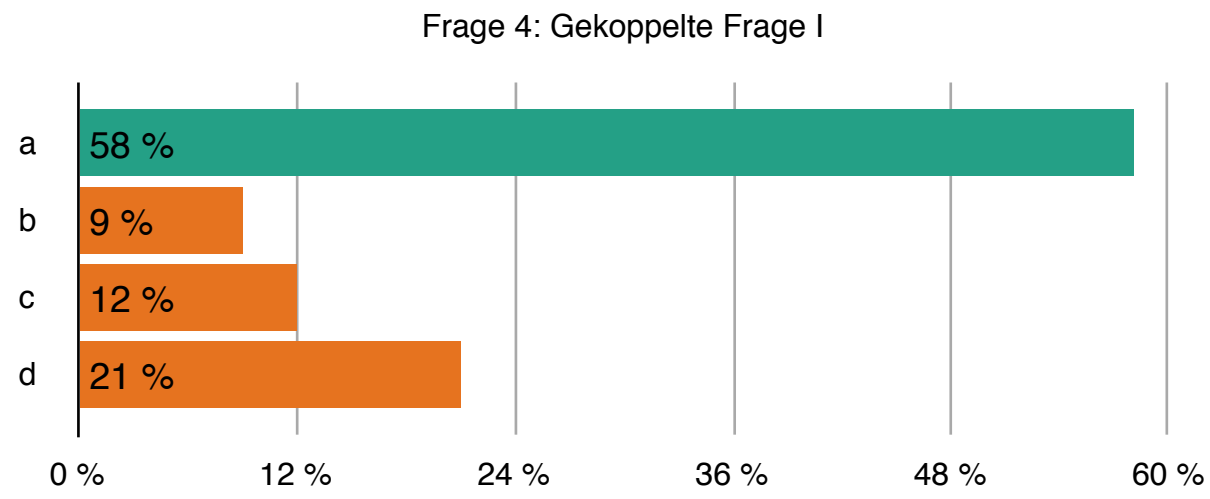
- ☐ ...kein Hemni stolk ist. (10)
- ☑ **...Goren meistens sehr flop sind. (15)**
- ☐ ...Goren immer stimpen. (17)
- ☐ ...alle Hemni dompeln. (10)

Auswertung:

Bei dieser Frage verhält sich die Mehrzahl der Testpersonen anders als vermutet. Nur 29% der Personen kreuzten die zu erwartende Antwort an, was knapp über der Lösungswahrscheinlichkeit von einem Viertel liegt. Die dritte, „falsche“ Option „Goren immer stimpen“ wurde sogar noch von 33% der Testpersonen angekreuzt. Dieses Ergebnis bestätigt nicht, dass qualifizierende Beiwörter wie „meistens“ dazu verleiten, diese Antwort als richtig zu empfinden.

Auffällig ist allerdings, dass die meisten Antworten auf die beiden Optionen entfielen, die sich auf das Subjekt des Satzes bezogen (insgesamt 62%). Es könnte weiter getestet werden, ob subjektbezogene Antworten eher angekreuzt werden als objektbezogene Antworten und unter welchen Umständen dies der Fall ist. Vermutlich werden in dieser Aufgabe die subjektbezogenen Antwortoptionen bevorzugt, da nach der Begründung für den Sieg der „Hemni“, die in der Funktion des Subjekts stehen, gefragt wird.

4.4 Frage 4: Gekoppelte Frage I



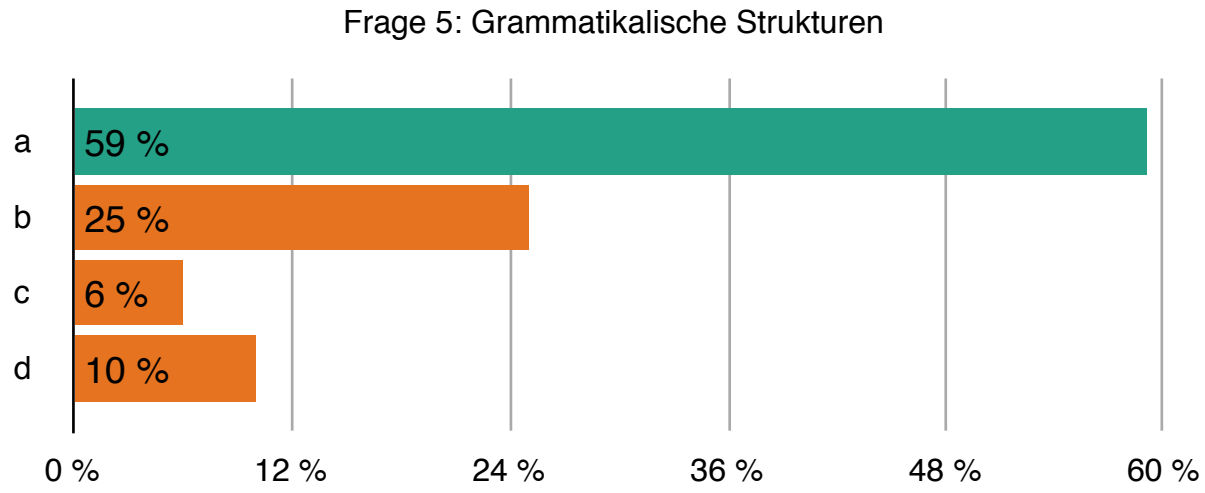
Womit wird der hesterinate Groller geöffnet?

- [x] **Einem Pflopf (30)**
- [ ] Einer Kierke (5)
- [ ] Einer Lierse (6)
- [ ] Einem Redel (11)

Auswertung:

Die erste der gekoppelten Fragen wird von 58% der Testpersonen den Vermutungen entsprechend beantwortet. Dies liegt weit über der Lösungswahrscheinlichkeit von einem Viertel. Die Testpersonen müssen sich intensiv mit dem Test beschäftigen, um diese Frage richtig beantworten zu können. Denn die Antwort für die Frage geht erst aus der siebten Frage hervor. Da diese Frage gleichzeitig auch die Antwort auf Frage 7 gibt, könnte man folgern, dass diese auch mindestens 58% Erfolgsquote verzeichnen müsste.

4.5 Frage 5: Grammatikalische Strukturen



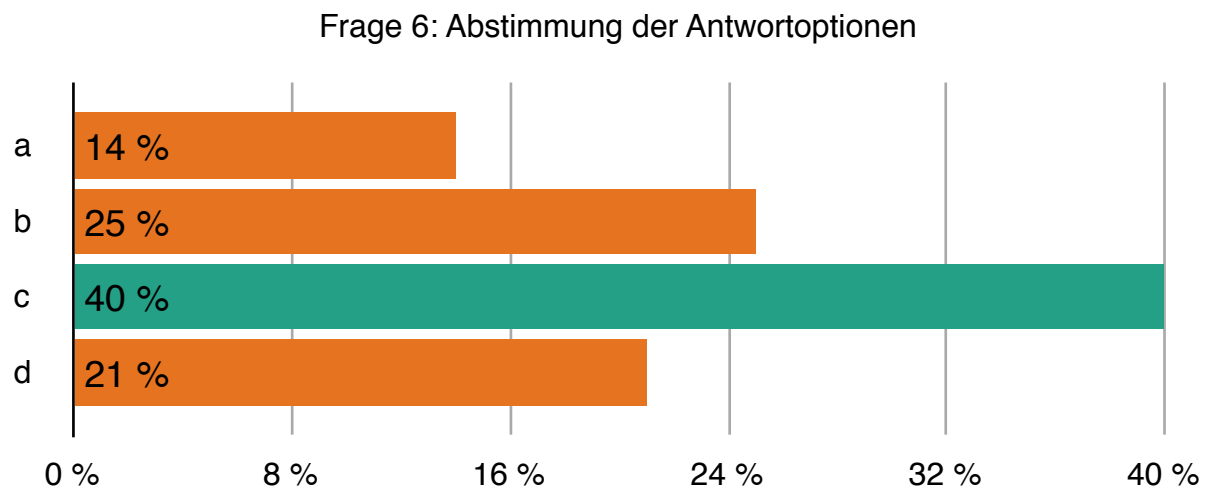
Die Ursachen für die Falkverwangung sind...

- [x] **...die Zottler und das Gasperatum. (31)**
- [ ] ...ein Holk im Geranium. (13)
- [ ] ...Hestina mit Verstinden spielen. (3)
- [ ] ...einige Walkos haben Trasper verloren. (5)

Auswertung:

Auch bei der fünften Frage ist ein deutliches Ergebnis von 59% Erfolgsquote zu verzeichnen. Vielen Testteilnehmer fallen somit die Verstöße gegen die deutsche Grammatik der Distraktoren auf, und durch das Ausschlussverfahren können sie auf die provozierte Antwort schließen. Allerdings ist es auffällig, dass eine recht große Anzahl der Testpersonen diese Frage mit anderen Optionen beantworten, obwohl bei genauem Lesen deutlich sein müsste, dass die Grammatik der Distraktoren nicht korrekt ist. Ich vermute, dass dies auf oberflächlichen Kontakt mit den Aufgaben schließen lässt und eine gewichtigere Testsituation die „Erfolgs“quote noch steigern würde.

4.6 Frage 6: Abstimmung der Antwortoptionen



Welche Testona gibt es **nicht**?

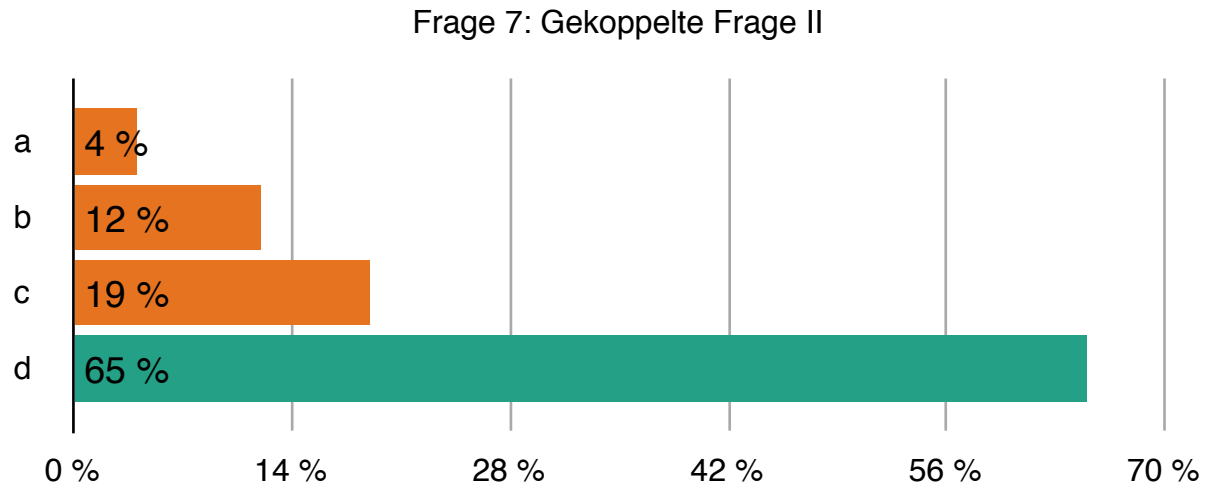
- ☐ Garst und Leonta (7)
- ☐ Flocker und Leonta (13)
- ☑ **Leonta (21)**
- ☐ Plosper und Leonta (11)

Auswertung:

Diese Frage ist wahrscheinlich diejenige, die am wenigsten von den Kunstwörtern beeinflusst wird. Lediglich logisches Denken führt zur provozierten Antwort. Denn „Leonta“ muss richtig sein, da es in allen Antworten vorkommt.

Dieses Muster haben 40% der Testpersonen durchschaut. Somit kann bestätigt werden, dass viele Testpersonen wahrnehmen, wenn sich die Antwortoptionen untereinander beeinflussen, und dies als Hilfestellung betrachten.

4.7 Frage 7: Gekoppelte Frage II



Die wichtigste Funktion eines Pflopfes kann nur in Verbindung mit welchem Gegenstand ausgeübt werden?

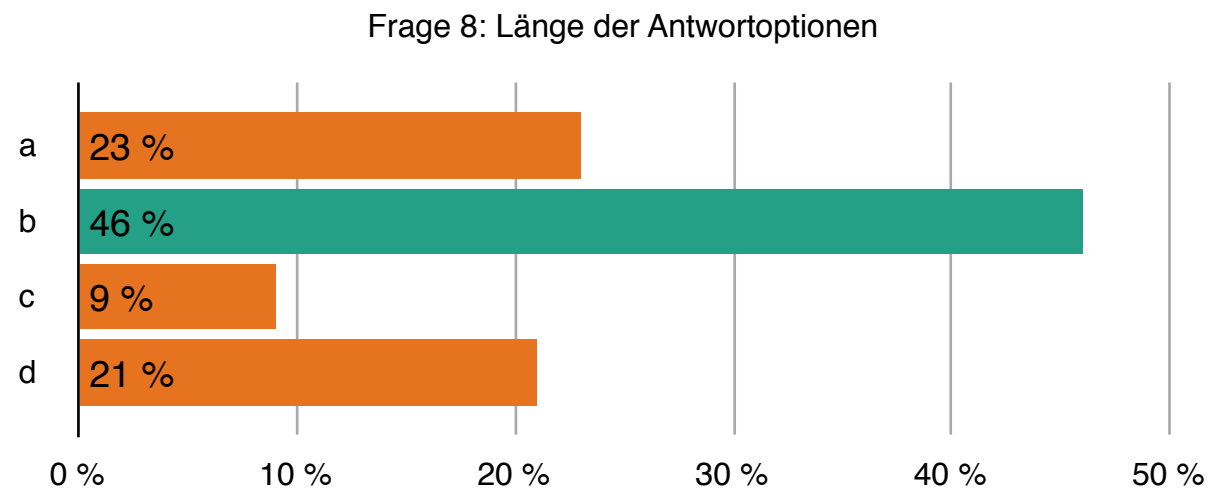
- ☐ Mit der moleren Quaterne (2)
- ☐ Mit einem exinen Pavanus (6)
- ☐ Mit dem Ropf eines Fertuns (10)
- ☑ **Mit einem hesterinatem Groller (34)**

Auswertung:

Die zweite der gekoppelten Fragen hat noch eine um 7% höhere Erfolgsquote als Frage Nr. 4. Diese Differenz zur vorigen Frage kann durch den Reihenfolgeneffekt entstanden sein. Die Testpersonen lasen das Wort „hesterinater Groller“ bereits in der vorangegangenen Aufgabe 4 und konnten diese Frage dadurch richtig beantworten. Jedoch versuchten sie im Gegenzug nicht, die erste der beiden Fragen durch die zweite zu beantworten. Es ist auch denkbar, dass sich die Probanden angesichts der geringen Bedeutung des Tests nicht die Mühe machten, noch einmal zurückzublättern und ihre Antwort auf Frage 4 zu korrigieren.

Eventuell wird aber auch das Wort „hesterinater Groller“ aufgrund seiner Länge besser memoriert, als das Wort „Pflopf“. Dadurch wäre eine höhere „Erfolgs“quote dieser Aufgabe zu erklären.

4.8 Frage 8: Länge der Antwortoptionen



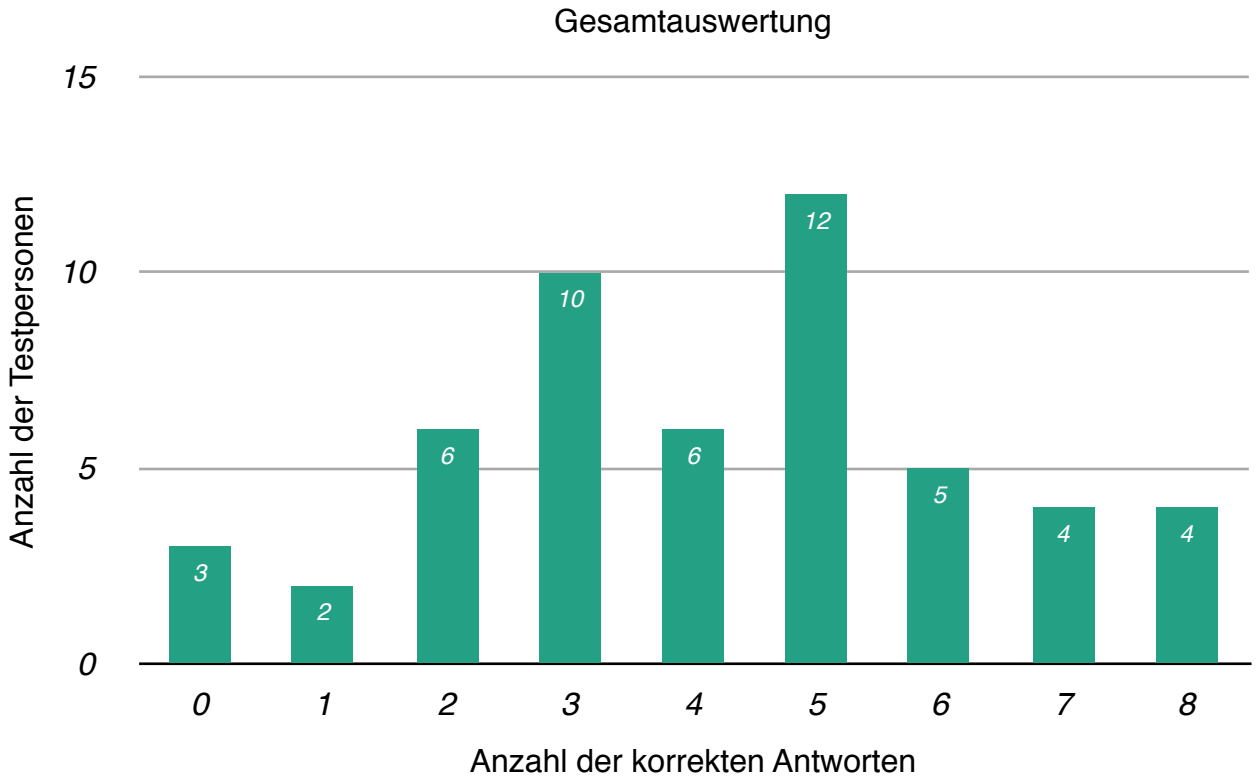
Warum ist der Barster fruppig?

- ☐ Weil er lange gemmert. (12)
- ☑ **Weil er nicht reddern kann und mehr Hopfer macht, wenn er gasser ist und kermig floppt. (24)**
- ☐ Weil er sehr doddelt ist. (5)
- ☐ Weil er seine Plocker maut. (11)

Auswertung:

Die achte Frage kann nur schwer durchschaut werden, da hier lediglich die Länge der Antworten variiert wird, ansonsten aber keinerlei Bezug auf diese oder andere Fragen genommen wird. Dennoch haben 46% der Testteilnehmer die provozierte Antwortoption angekreuzt. Somit ist bestätigt, dass längere Antworten den Testteilnehmer dazu verleiten, sie als richtig anzusehen. Daher sollten in MC-Tests die Antwortoptionen möglichst in der gleichen Länge verfasst sein.

4.9 Gesamtauswertung



Durchschnittlich haben die Testpersonen vier der acht MC-Fragen den Erwartungen entsprechend beantwortet. Dies entspricht einer Quote von 50% und liegt somit weit über der Lösungswahrscheinlichkeit von 25%. Die Testpersonen haben viele der MC-Fragestrukturen durchschaut und konnten sie nur deshalb richtig beantworten. Vier Personen (8%) lösten sogar alle Fragen richtig“. Die Personen, die keine oder weniger als vier Aufgaben richtig gelöst haben, schrieben meist zu ihrer Beantwortung des Fragebogens, dass sie die Antworten ankreuzten, die „gut klangen“. Daher gehe ich davon aus, dass sie den Fragebogen intuitiv beantworteten und nicht versuchten, Strukturen zu durchschauen.

Auffällig ist auch, dass jede einzelne Frage überzufällig häufig im Sinne der im Itemstamm provozierten Antwortoption beantwortet wurde. Dies führt mich zu der Annahme, dass die Strukturen der Fragen die Antwort der Testteilnehmer stark beeinflussen. Lediglich bei der Beantwortung der dritten Frage präferierten mehr Testpersonen einen Distraktor als Antwort. Dennoch haben auch hier mehr Testpersonen als der Lösungswahrscheinlichkeit entsprechen (29%) die Frage entsprechend der provozierten Antwort gelöst. Um zu testen, ob dies wirklich auf der Struktur und Wortwahl der Frage beruht, würde ich die These der Frage erneut in einem separaten, umfangreicherem Fragebogen testen.

Insgesamt folgere ich aus den Ergebnissen, dass die durch den Fragebogen getesteten möglichen Fehler bei den Testpersonen starken Einfluss auf die Wahl der Antwortoptionen haben. Daher sollte jeder Testentwickler genau auf die Formulierung seines Itemstamms und der Antwortoptionen achten und die in diesem Fragebogen aufgezeigten Fehler umgehen, um das tatsächliche Leistungsvermögen einer Testperson bestimmen zu können.

5. Fazit

Zweck der vorliegenden Arbeit war ein doppelter: Zum einen sollte eine deutsche Version des amerikanischen Fragebogens „Test your testwiseness“ erstellt werden. Dieses Material basiert auf der Verwendung von Kunstwörtern, sodass es offensichtlich unmöglich ist, auf Basis von Vorwissen korrekte Antworten zu geben. Testpersonen, die dennoch eine Antwortoption wählen sollen, müssen ihre Entscheidung auf Basis anderer Überlegungen treffen. Dieses Material erweist sich als überaus instruktiv für die Verdeutlichung von typischen Fehlern bei der Itemkonstruktion. Dieser Effekt geht aber verloren, wenn die TeilnehmerInnen wegen fehlender Sprachkenntnisse die Kunstwörter nicht als solche identifizieren können, sondern sie für „echte“ englische Wörter halten. Mit dieser Arbeit liegt nun eine analoge deutsche Fassung des Tests vor.

Zum anderen sollte an dem konstruierten Testmaterial gezeigt werden, dass die in der Literatur beschriebenen typischen Fehler bei der Itemkonstruktion tatsächlich Auswirkungen auf das Ankreuzverhalten von Probanden haben und damit die Objektivität der Items verletzen. Zu diesem Zweck wurden die Testitems gezielt hinsichtlich jeweils eines bestimmten Konstruktionsfehlers manipuliert und einer größeren Gruppe von Testpersonen vorgelegt. Es zeigte sich, dass die Probanden in jedem Fall überzufällig häufig die vorhergesagte Option wählten. Damit ist die Wirksamkeit der Konstruktionsfehler auch empirisch nachgewiesen und das Instrument kann - trotz der reduzierten Durchführungsobjektivität der Studie - als empirisch erprobt gelten.

Literaturverzeichnis

- Asmuth, Markus (2003): Prüfen mit der Multiple-Choice-Methode. online verfügbar unter: http://www.lehrer-online.de/multiple-choice.php?sid=354199401943769644259001000107 90 [Zugriff am 09.03.2010]
- Bremerich-Vos, Albert/ Köller, Olaf/ Behrens, Ulrike/ Bühme, Katrin/ Dietrich, Fabian/ Granzer, Dietlinde/ Krelle, Michael/ Neumann, Daniela/ Robitzsch, Alexander/ Winkelmann, Henrik (2007): Hinweise zur Erstellung von Testaufgaben für das Projekt „Evaluation der Standards Deutsch für die Sekundarstufe I" (ESDeS). Unveröffentlichtes Arbeitsmaterial.
- ETH Zürich (Hg.) (o.J.): Wegweiser für gute Multiple-Choice-Fragen. online verfügbar unter: http://www.elba.ethz.ch/docs/mcfragen.pdf [Zugriff am 13.03.2010]
- Gronlund, Norman E. (2006): Assessment of Student Achievement. 8. Auflage. Boston: Pearson Education.
- Jacobs, Bernhard (2002): Richtlinien von einfachen Multiple-Choice-Aufgaben nach Gronlund. online verfügbar unter: http://www.phil.uni-sb.de/FR/Medienzentrum/verweise/psych/aufgaben/mcguideline.html [Zugriff am 08.03.2010]
- Leibniz Universität Hannover (Hg.) (o.J.): Erstellen und Bewerten von Multiple-Choice-Aufgaben. online verfügbar unter: http://www.uni-hannover.de/imperia/md/content/elearning/practicalguides2/didaktik/elsa_handreichung_zum_erstellen_und_bewerten_von_mc_fragen.pdf [Zugriff am 13.03.2010]
- Lienert, Gustav/ Raatz, Ulrich (1994): Testaufbau und Testanalyse. 5. Auflage. Weinheim: Beltz, Psychologie-Verlags-Union.
- New York State Education Department (Hg.) (o.J.): Test your Testwiseness. online verfügbar unter: http://www.emsc.nysed.gov/osa/assesspubs/pubsarch/ActivityTestYourTestwiseness.pdf [Zugriff am 08.03.2010]
- Schmidts, Michael/ Lischka, Martin (2001): Prüfungsfragen für Multiple-Choice Tests erstellen. Kurzanleitung mit Beispielen. online verfügbar unter: http://www.med.uni-giessen.de/intranet/lehre/Anleitung_Erstellung_von_MC-Fragen.pdf [Zugriff am 10.03.2010]