



universität
wien

DIPLOMARBEIT

Titel der Diplomarbeit

„Feedback von Schülern: subjektiv wahrgenommene
Eigenschaften von Multiple Choice Tests in
IT - Fächern“

Verfasserin

Claudia Hladecek

angestrebter akademischer Grad

Magistra der Philosophie (Mag. phil.)

Wien, im September 2009

Studienkennzahl lt. Studienblatt:
Studienrichtung lt. Studienblatt:
Betreuer:

A 297
Pädagogik
Univ.-Prof. Dr. Stefan Hopmann

DANKSAGUNG

Ein herzliches Dankeschön geht an Univ.-Prof. Dr. Stefan HOPMANN, der mich lehrreich und hilfsbereit während der ganzen Diplomarbeit betreut hat. An dieser Stelle möchte ich mich vor allem auch bei meiner Familie für ihre liebevolle Unterstützung bedanken, ohne deren Hilfe diese Arbeit nicht möglich gewesen wäre. Mein Dank gilt auch Dipl.-Ing. Dr. Walter FÜHRER, der mir seinen Datensatz zur Verfügung gestellt hat und mich stets beratschlagt hat. Ich möchte mich auch bei meinen Freunden und Studienkollegen bedanken, die mich immer wieder motiviert haben und mir mit Rat und Tat zur Seite standen.

Kurzfassung

Diese Diplomarbeit beschäftigte sich mit der Frage, ob sich Unterschiede in der Einschätzung vom Multiple Choice Antwortformat gegenüber offenen Antwortformaten bei Schülern feststellen lassen.

Die Literaturrecherche zeigte, dass – obwohl zum Thema Multiple Choice Tests bereits etliche Untersuchungen durchgeführt wurden – noch sehr wenig Evidenz zur Einschätzung von Schülern dieser Fragen vorliegt. Daher wurde in dieser Arbeit eine Längsschnittsstudie über drei Schuljahre durchgeführt, bei der insgesamt 280 Schüler teilnahmen. 24 Schüler wurden dabei über alle drei Jahre getestet. Als Erhebungsmethode wurde ein Feedbackbogen gewählt, der Fragen über einen vorangegangenen Multiple Choice Test beinhaltete. Die Ergebnisse dieser Untersuchung gaben Hinweise darauf, dass Schüler die bei dem Multiple Choice Test gut abschnitten diesen als einfacher und den Lernaufwand als niedriger bewerteten, als Tests mit offenen Fragen. Es konnte jedoch nicht nachgewiesen werden, dass Schüler ein Prüfungsformat bevorzugen, noch dass sich das Antwortverhalten über die Testzeitpunkte wesentlich verändert.

Abstract

The present masters' thesis investigates whether or not there are differences between students' ratings of multiple-choice and open-ended test questions.

The literature review showed that although there already had been many studies on the question of multiple-choice tests, there was very little on how students saw such tests. Therefore, the present study employed a longitudinal design, comprising data on a total of 280 students over a period of three years, whereby 24 students were tested over each of the three years. The method used was a feedback sheet with questions on a previously completed multiple-choice test. Results of this study indicated that students who did well at the multiple-choice tests found these to be simpler and considered the related learning effort to be of lesser value than that needed for open-ended questions. However, there was no evidence that students preferred one type of test format or that attitudes towards these formats changed during the testing period.

Inhaltsverzeichnis

I.	Einleitung.....	3
II.	Theoretischer Teil.....	5
II.1	Prüfungsformen.....	5
II.2	Exkurs: Leistungsbeurteilung.....	8
II.2.1	Leistungsbewertung.....	8
II.2.2	Funktion von Noten.....	9
II.2.3	Leistungsmotivation	10
II.2.4	Zusammenfassung: Leistungsbeurteilung.....	11
II.3	Aufgabenformen.....	13
II.3.1	Schwierigkeiten.....	14
II.3.2	Offene Aufgaben.....	15
II.3.3	Halb offene Aufgaben	15
II.3.4	Geschlossene Aufgaben	16
II.3.5	Zusammenfassung: Aufgabenformen	17
II.4	Offene Fragen	19
II.4.1	Vorteile von offenen Fragen.....	19
II.4.2	Nachteile von offenen Fragen	20
II.4.3	Zusammenfassung: Offene Fragen.....	21
II.5	Multiple Choice Fragen.....	22
II.5.1	Multiple Choice Tests.....	23
II.5.2	Vorteile von Multiple Choice Fragen.....	23
II.5.3	Nachteile von Multiple Choice Fragen.....	25
II.5.4	Ratewahrscheinlichkeit	26
II.5.5	Zusammenfassung: Multiple Choice Fragen	27
II.6	Stand der Forschung.....	28
II.6.1	Multiple Choice versus Offene Fragen	28
II.6.2	Einschätzung und Präferenz	30
II.6.3	Speziell zum Thema Multiple Choice Tests.....	35
II.7	Schülerurteile	40
II.8	Zusammenfassung.....	42
III.	Empirischer Teil	43
III.1	Methode	43
III.1.1	Erhebungsinstrumente	43
III.1.2	Datenerhebung	46
III.2	Auswertung der Ergebnisse.....	47
III.2.1	Stichprobenbeschreibung	47
III.2.2	Datenaufbereitung	48
III.3	Darstellung der Ergebnisse	51
III.3.1	Korrelationen	51
III.3.2	Allgemeines lineares Modell für Messwiederholungen	65

III.3.3 t-Test bei abhängigen Stichproben	68
III.4 Beschreibung der Ergebnisse	71
III.4.1 Zusammenhang zwischen der Testleistung und den Feedbackfragen.....	71
III.4.2 Veränderungen über die Zeit, bei mehr als zwei Testzeitpunkten	72
III.4.3 Veränderungen über die Zeit, bei zwei Testzeitpunkten	74
III.5 Interpretation.....	76
III.6 Diskussion	79
III.7 Resümee und Ausblick	80
Appendix A.....	83
Appendix B.....	88
Literaturverzeichnis	89
Lebenslauf.....	96

I. Einleitung

Im Zuge der Literaturrecherche zum Thema Multiple Choice Tests und meiner eigenen Erfahrungen in meiner Schulzeit konnte ich feststellen, dass Multiple Choice Aufgaben im österreichischen Schulwesen, ein unübliches Aufgabenformat darstellen (PESCHEK, SCHNEIDER 2006, 77). In Deutschland und Österreich gehören offene Aufgaben zu den gewohnten Aufgabenformaten in der Schule (BAUMERT, LEHMANN 1997, 49f). An den Universitäten hingegen wird diese Form des Fragens immer häufiger eingesetzt. In Amerika sind Multiple Choice Tests sehr populär und werden unter anderem für die Zulassungsprüfungen auf Hochschulen eingesetzt (BAR-HILLEL, u.a. 2005, 4). Aufgrund dieser Umstände beschäftigte ich mich mit der Frage, inwiefern Schüler¹ einer Höheren Technischen Lehranstalt Multiple Choice Tests einschätzen beziehungsweise wahrnehmen. Dazu wurde den Schülern ein Feedbackbogen, nach Vollendung eines Multiple Choice Tests, vorgelegt. Neben Fragen zu diesem Prüfungsformat beinhaltete dieser Fragebogen, Fragen zu Prüfungen mit offenen Fragen. Unter offenen Fragen wurden jene verstanden, die keine vorgefertigten Antworten vorgaben.

Vorweg sei angemerkt, dass es nicht mein Anliegen war herauszufinden, ob Multiple Choice Tests besser oder schlechter als Tests mit offenen Fragen sind, vielmehr ging es mir darum festzustellen, wie Schüler Multiple Choice Tests bewerten und einschätzen.

Im Theoretischen Teil dieser Arbeit werden im ersten Kapitel zunächst einmal die verschiedenen Arten von Prüfungsformen vorgestellt, die der Feststellung und Bewertung schulischer Leistung dienen. Es wird weiters ein kurzer Einblick in die Funktion von Noten und die Leistungsmotivation gegeben, da diese wichtige Kriterien für die Erhebung von Prüfungen darstellen. Im zweiten Kapitel werden

¹ Die verwendeten Personenbezeichnungen beziehen sich auf beide Geschlechter, männlich, weiblich, wobei ich jedoch zur besseren Lesbarkeit der Arbeit die männliche Form verwende.

die unterschiedlichsten Aufgabenformen schriftlicher Prüfungen vorgestellt, um dann in den zwei darauf folgenden Abschnitten näher auf offene Fragen und Multiple Choice Fragen einzugehen.

In der Literatur findet man oft Argumente gegen Multiple Choice Tests. Unter anderem, weil diese Aufgaben teilweise durch raten gelöst werden können, sie nur das Wiedererkennen von Gedächtnisinhalten verlangen, man mit diesem Aufgabentyp nur Faktenwissen und kein Verständnis von Zusammenhängen prüfen kann. Diesen Aufgaben werden oft Aufsätze mit offenen Fragen gegenüber gestellt, denen in der Literatur häufig nachgesagt wird, sie würden selbstständiges Denken, Problemlösen und Verständnis erfassen (INGENKAMP 1989, 180).

Aufgrund meiner Fragestellung: „Wie schätzen Schüler Multiple Choice Fragen im Gegensatz zu offenen Fragen ein?“ wird im Theorieteil noch der Stand der Forschung, zum Thema Multiple Choice Tests, sowie die Vor- und Nachteile von Schülerurteilen, näher erläutert.

II. Theoretischer Teil

In diesem Teil der Arbeit wird ein Überblick über die Prüfungsformen und den damit verbundenen Aufgabentypen gegeben. Besonderes Augenmerk wird dabei auf schriftliche Prüfungen mit offenen Fragen und Multiple Choice Fragen gelegt. Weiters wird eine Einführung in den „Stand der Forschung“ im Bereich von Multiple Choice Tests, unter Berücksichtigung der vorangegangenen Erkenntnisse, vorgestellt und beschrieben und die Thematik der Schülerurteile kurz erläutert.

II.1 Prüfungsformen

Prüfungen, die im Unterricht stattfinden, werden als Fremdprüfungen im weiteren Sinn bezeichnet. Fremdprüfungen im engeren Sinn sind jene, bei denen sich Prüfer und Prüfling nicht, beziehungsweise nur kaum kennen (SCHWENDENWEIN 1985, 150). In dieser Arbeit wird nur Bezug auf Prüfungen im weiteren Sinn genommen.

Die Wahl der Prüfungsform hängt nach SACHER (2004, 60) von der Funktion der Leistungsüberprüfung ab. „Schriftliche Überprüfungen umfassen ein in sich abgeschlossenes kleineres Stoffgebiet. Folgende Formen schriftlicher Überprüfungen sind zulässig: a) Tests, b) Diktate [...]“. (LBVO §8 (1))

Schriftliche Prüfungen sind im Gegensatz zu mündlichen Prüfungen nicht nur ökonomischer durchzuführen, sie liefern auch vergleichbare Ergebnisse (SACHER 2004, 60). Ein zusätzlicher Vorteil dieser Prüfungsform liegt darin, dass das Produkt der Prüfung beständig ist, somit stets wiederholt betrachtet werden kann. Bei mündlichen Prüfungen hingegen verblasst die Leistung schnell. Weiters sind schriftliche Prüfungen oft objektiver und reliabler. Im Gegensatz dazu, liegt der Vorteil von mündlichen Prüfungen darin, dass der Prüfer Missverständnisse

aufgrund der Fragestellung beheben kann und der Prüfling die Möglichkeit hat nachzufragen (CZERWENKA 1990, 28).

Objektiv sind Prüfungen dann, wenn ihre Ergebnisse unabhängig vom Prüfer sind. Dabei kann zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität unterschieden werden. Bei der ersten Art, geht es darum, ob auch andere die Prüfung so gestalten würden, bei der Zweiten, ob auch andere Prüfer die Ergebnisse so auswerten würden und bei der Letzten, ob verschiedene Personen die Ergebnisse gleich bewerten würden (SACHER 2004, 35f).

Reliabilität meint die Zuverlässigkeit, also Genauigkeit und Sicherheit einer Messung. Die Überprüfung der Reliabilität erweist sich unter schulischen Bedingungen als sehr schwierig, da weder eine Wiederholung des gleichen Tests, die Trennung des Tests in zwei strukturgleiche Hälften noch zwei gleiche Tests nacheinander durchgeführt werden können (SACHER 2004, 36f). Um die Reliabilität zu bestimmen, müssten Lehrer Schularbeiten oder Tests nach mehreren Wochen gleich benoten beziehungsweise zu der gleichen Punktezahl kommen, was jedoch in der Praxis selten gemacht wird, da die Zeit dazu fehlt (BESSOTH 2002, 31). Je weniger Aufgaben eine Prüfung hat, desto geringer ist die Reliabilität (SACHER 2004, 44).

Die Gültigkeit einer Messung wird Validität genannt. Um die Validität von Leistungsbeurteilungen zu verbessern, stehen dem Lehrer neben den verschiedenen Prüfungsformen – mündlich und schriftlich – den Prüfungssituationen – in der Klasse oder allein – auch die Aufgabenformen – offen oder geschlossen – zur Verfügung (SACHER 2004, 37ff).

Der Einsatz von schriftlichen Prüfungen ist bei der Befragung von vielen Schülern und bei Zeitmangel sinnvoll. Ist die Reihenfolge der Aufgaben unwichtig, die gleichen Prüfungsbedingungen erwünscht und der Kontakt zum Prüfer unerwünscht, bietet sich ebenfalls diese Prüfungsform an (SCHWENDENWEIN 2000, 397). Bei mündlichen Prüfungen ist ein enger Kontakt zum Prüfer gegeben und die Prüfungsbedingungen – die Prüfungsleistung vom ersten Schüler beeinflusst die Bewertung des zweiten – sind nicht für alle Schüler gleich (SCHWENDENWEIN 2000, 395).

SACHER (2004, 58) vertritt die Meinung, dass in vielen Fächern auf verschiedene Weise geprüft werden kann. Die Entscheidung über die Prüfungsform bleibt dem Lehrer überlassen.

„Die vom Lehrer jeweils gewählte Form der Leistungsfeststellung ist dem Alter und dem Bildungsstand der Schüler, den Erfordernissen des Unterrichtsgegenstandes, den Anforderungen des Lehrplanes und dem jeweiligen Stand des Unterrichtes anzupassen.“ (LBVO §2 (3))

Da „in Prüfungen [...] die Reaktion auf Anforderungen nach bestimmten Kriterien beurteilt oder bewertet“ wird (KLINGLER 1982, 31), gebe ich nun einen kurzen Einblick in die Leistungsbeurteilung und -bewertung.

II.2 Exkurs: Leistungsbeurteilung

Bis in das hohe Mittelalter war Bildung den Adeligen vorbehalten, bis zu dem Zeitpunkt, wo diese erkannten, dass „bürgerliche Beamte machtpolitische Vorteile für sie bringen“. Dies führte Ende des 19. Jahrhunderts dazu, dass die Beurteilung durch die Lehrer aufgrund sozialer Mobilität immer wichtiger wurde. Soziale Mobilität heißt in diesem Kontext, dass „begabte Söhne niederer Herkunft [...] weiterführende Schulen besuchen“ konnten. Im 20. Jahrhundert entwickelten sich die Schulen – standespolitische Kriterien fehlten und der Andrang wuchs – zu Leistungsschulen (LISSMANN 1998, 11). Dies führte in weiterer Folge dazu, dass das Erbringen von Leistung sowie die Bewältigung von Leistungsanforderungen zu den Kerngeschäften von Schulen wurde (HOSENFELD, SCHRADER 2006, 9) und darüber hinaus aufgrund gezeigter schulischer Leistungen weitere Bildungswege und Berufschancen mehr oder weniger eröffnet wurden (HAIDER, SCHREINER 2006, 229).

Da die Problematik über den Erfolg und das Versagen in der Schule immer mehr in die Öffentlichkeit rückt, werden die Methoden und Kriterien, mit denen in der Schule Schülerleistungen beurteilt werden, nicht nur pädagogisch, sondern auch politisch geprägt (PÖGGELER 1991, 216). Dies zeigt sich in den Diskussionen um die PISA Studie, wo aufgrund schlechter Leistungen politische Reformen beschlossen werden und dies sich in weiterer Folge auf die Prüfungs- und Unterrichtspraxis in der Schule auswirkt.

II.2.1 Leistungsbewertung

Der Ansatz der Bezugsnormorientierung konzentriert sich „vorrangig auf Leistungsbewertung und Leistungsrückmeldung in Abhängigkeit von Leistungsstand und Leistungsentwicklung der Schüler.“ (CLAUSEN 2002, 20)

SACHER (2004, 87) unterscheidet drei Bezugsnormen der Leistungsbeurteilung. Wird die Leistung eines Einzelnen im Verhältnis zu einer Gruppenleistung beurteilt, handelt es sich um die soziale Norm. Die individuelle Norm ist gekennzeichnet durch den Lernfortschritt eines Schülers. Bei der kriterialen Norm ist die Beurteilung unabhängig von der Gruppenleistung, sie bezieht sich auf fachlich-sachliche Anforderungen, wie zum Beispiel das angestrebte Lernziel (siehe auch JÜRGENS 2005, 46). „In der Praxis tritt meist eine Mischung zwischen den Bezugssystemen auf.“ (HANISCH 2003, 206)

Leistungsfeststellung kann nach HANISCH (2003, 206) durch Beobachtung über einen längeren Zeitraum oder punktuell, also schriftlich – durch eine Arbeit – oder mündlich stattfinden. Das Ergebnis der Leistungsfeststellung wird durch die Leistungsbeurteilung berichtet, entweder in Form von Noten oder verbalen Berichten (HANISCH 2003, 206). Da im österreichischen Schulwesen zur Zeit noch hauptsächlich Noten die Leistungen von Schülern widerspiegeln, wird hier nicht näher auf verbale Beurteilungen eingegangen.

II.2.2 Funktion von Noten

Durch die Einführung von Noten wollten schon die Jesuiten sichern, dass die Anstrengungen der Schüler wahrgenommen werden und deren Lernfreude, Wetteifer sowie Leistungen bestärken (LIEDTKE 1991, 25). Noten dienen der Leistungsbeurteilung und haben sowohl eine Rückmeldefunktion für den Lehrer, ob der Unterricht erfolgreich war, als auch eine Rückmeldefunktion für den Schüler, um dessen eigene Leistung zu veranschaulichen (ROLLETT 1997, 171).

„Für die Beurteilung der Leistungen der Schüler sind folgende Beurteilungsstufen (Noten) zu verwenden: Sehr Gut (1), Gut (2), Befriedigend (3), Genügend (4), Nicht genügend (5).“ (SchUG §18 (2)) „Durch die Noten ist die Selbständigkeit der Arbeit, die Erfassung und die Anwendung des Lehrstoffes, die Durchführung der Aufgaben und die Eigenständigkeit des Schülers zu beurteilen.“ (SchUG §18 (3))

Nach BESSOTH (2002, 19) sind Noten weitgehend leistungsfreie Aussagen, da sie lediglich über den Stand eines Schülers in einer bestimmten Lerngruppe informieren. GAGE und BERLINER (1996, 655) sind der Ansicht, dass „Noten helfen, die Stärken und Schwächen eines Schülers zu identifizieren.“

Bei einer Untersuchung von JACHMANN und TILLMANN (2000, 42) konnte beobachtet werden, dass die Leistungsbeurteilung, im Speziellen die eindeutige Positionierung durch Noten, sowohl von Schülern, Eltern und Lehrern mehrheitlich gewünscht wurde.

II.2.3 Leistungsmotivation

„Leistungsüberprüfungen und -beurteilungen haben Rückwirkungen auf die Leistungserziehung der Schüler und sind Teil einer Erziehung der Schüler zur eigenverantwortlichen Gestaltung ihrer Lernprozesse.“ (SACHER 2004, 30) Leistungserziehung ist notwendig, da auf der einen Seite Lehrer Lernprozesse nur anregen und unterstützen, auf der anderen Seite der Schüler seine Lernprozesse nach der Schulzeit ohne fremde Hilfe organisieren können muss. Autonome Lernhandlungen müssen beigebracht werden, damit der Schüler seine Lernprozesse selbst steuern und verantworten kann (SACHER 2004, 30). Denn bei der Erziehung zur Lern- und Leistungsbereitschaft geht es nicht primär um den Nachweis von Wissen und Können, sondern vielmehr um „das Erlernen der Fähigkeit, Wissen und Können selbständig und eigenverantwortlich in sozialen, emotionalen und kognitiven Handlungs- und Lernvollzügen anzuwenden und weiterzuentwickeln“ (JÜRGENS 2005, 33). Dabei spielt die Leistungsmotivation eine große Rolle.

Leistungsmotivation ist „das Streben des einzelnen nach Erfolg, danach, etwas gut zu können.“ (GAGE, BERLINER 1996, 344) Die erbrachte Leistung stellt einen Indikator für Leistungsmotivation dar. Erfolg und Misserfolg werden Ursachen, so genannte Attributionen, zugeschrieben, die das nachfolgende Verhalten beeinflussen. Bei der Wahrnehmung von Ursachen handelt es sich nach WEINER

(1984, 221) „um eine Zuschreibung durch den Beobachter [...]; Ursachen *per se* sind nicht direkt beobachtbar“. Attributionen werden in Ursachenlokalisierung, Stabilität und Kontrollierbarkeit unterteilt (WEINER 1984, 295). Bei der Ursachenlokalisierung wird zwischen internen und externen Faktoren differenziert. Ein interner Grund für Erfolg wäre: Man ist gut in dem Fach; im Gegensatz dazu wäre eine externe Ursache zum Beispiel das Bestehen einer Prüfung aufgrund der Einfachheit der Aufgaben. Die Stabilität wird in stabile und instabile Faktoren unterteilt. Eine stabile Attribution würde sich zum Beispiel an eine Diskriminierung seitens des Prüfers binden, eine instabile dagegen an die eigene Anstrengung. Die Kontrollierbarkeit besteht aus kontrollierbaren und unkontrollierbaren Faktoren. Eine kontrollierbare Ursache ist zum Beispiel die Tatsache, nicht genug gelernt zu haben, eine unkontrollierbare könnte der Schwierigkeitsgrad der Prüfung sein. Gründe die von Schülern genannt werden, um eine Prüfung zu bestehen oder nicht, sind unter anderem die eigene Anstrengung, der Schwierigkeitsgrad der Aufgaben und Pech beziehungsweise Glück (GAGE, BERLINER 1996, 348ff; WEINER 1984, 258ff); konkret gesagt, ob sich der Schüler für seinen Erfolg oder Misserfolg selbst verantwortlich macht oder zum Beispiel den Lehrer.

Tests und Noten können eine motivierende Wirkung auf Schüler haben, die einen Wert in den Noten erkennen, was sich wiederum lernfördernd auswirkt (GAGE, BERLINER 1996, 381). Ist der Schüler der Ansicht, er habe keine Kontrolle über seinen Misserfolg, kann diese zu einem Abbruch seines Bemühens führen und der Lerneifer sinkt (GAGE, BERLINER 1996, 350).

II.2.4 Zusammenfassung: Leistungsbeurteilung

„Das Lernleistungsergebnis präsentiert sich letztlich in einer Prüfungssituation, wo das *Lernergebnis als aktueller Grad der individuellen Erreichbarkeit und des Könnens* (und nicht des Kennens) wahrgenommen werden kann.“ (SCHWENDENWEIN 2000, 358) Neben der Motivation des Schülers und den

Bewertungskriterien des Lehrers können die Leistungen der Klasse Auswirkungen auf die individuelle Erreichbarkeit haben (BOSSONG 2006, 263). Denn je nach Leistungsstand der Klasse werden den Schülern in der Regel Noten gegeben.

Zusammenfassend ist noch zu erwähnen, dass auch gesellschaftliche Entwicklungen einen Einfluss auf die Leistungsbewertung haben und in weiterer Folge auch auf die Prüfungen in der Schule. Es konnte gezeigt werden, dass es nicht nur allein in der Hand des Schülers liegt, wie gut er bei einem Test abschneidet. Dies wird in den nachfolgenden Kapiteln noch öfters aufgezeigt werden.

Abschließend werden noch die Begriffe Lehr- und Lernziel definiert, da diese im Verlauf dieser Arbeit immer wieder auftreten. „Die Lehrziele, die der oder die **Lehrende** formuliert, stellen für die Schüler und Schülerinnen **Lernziele** dar. Lehr- und Lernziel sollten einander möglichst genau entsprechen.“ (ROLLETT 1997, 158)

II.3 Aufgabenformen

Nach Festlegung der Prüfungsform, also schriftlich oder mündlich, kann eine Auswahl über die Aufgabenform getroffen werden (SACHER 2004, 61).

Eine Aufgabe besteht aus dem Informationsfeld, dem Fragefeld und dem Antwortfeld. Beim Ersteren handelt es sich um die Beschreibung von Ausgangsbedingungen, die man zur Lösungsfindung benötigt. Beim Fragefeld werden Probleme, die zu lösen sind, dargelegt. Beim Antwortfeld werden entweder Antworten vorgegeben, wie zum Beispiel bei Multiple Choice Fragen, oder es werden keine Antworten vorgegeben, wie zum Beispiel bei offenen Fragen (SACHER 2004, 61).

„Das Antwortfeld kann [...] sowohl der Lehrkraft als auch den Schülern zugänglich sein.“ (SACHER 2004, 62) Ist dem Schüler das Antwortfeld zugänglich, handelt es sich um geschlossene Aufgaben. Sind hingegen keine Lösungsalternativen vorgegeben, handelt es sich um halb offene Aufgaben, bei denen der Lehrer für sich die Antworten festlegt. Offene Aufgaben haben keine fix festgesetzten Antworten (SACHER 2004, 62).

Aufgaben werden aufgrund des jeweiligen Lehrziels formuliert und können wie bereits erwähnt unterschiedlichste Form annehmen. So unterscheidet man neben „Richtig/Falsch-“ und Multiple Choice- Antworten jene, bei denen frei formulierte Stellungnahmen oder konkrete Darbietungen verlangt werden. Letztere werden zum Beispiel beim Turnen zur Feststellung des Lehrziels herangezogen (HELLER, HANY 2001, 96). Die Antworten hängen vom Bezugsrahmen, vom Informationsstand des Befragten, von der Art und Positionierung der Frage und von der Erhebungssituation ab (FRIEDRICHS 1980, 193). Je nachdem was ein Lehrer vom Schüler erfragen will, wählt er die eine oder andere Aufgabenform. Dies hört sich vorerst nicht so kompliziert an, jedoch können Schwierigkeiten bei der Konstruktion von Aufgaben auftreten, die nun näher beschrieben werden. Denn schon SACHER (2004, 65) ist der Ansicht, dass das eigentliche Problem die Formulierung der Aufgaben vielen Lehrern Schwierigkeiten bereitet.

II.3.1 Schwierigkeiten

Nach PÖGGELER (1991, 217) sind unklar formulierte Aufgaben – gedanklich oder sprachlich – dem Lehrer anzulasten, da dieser die Aufgaben unter Umständen nicht altersadäquat stellt. MEMMERT (1997, 273) schreibt in seinem Artikel „Kunstfehler beim Unterrichten“: „Vor dem Stellen einer Aufgabe muß eine Aufgabenanalyse erfolgen nach Parametern wie umfangreich/kurz, mündlich/schriftlich, präzise/vage, leicht/schwer, eng/weit, allein/in Gruppen/mit Partner, benötigte Hilfsmittel u.ä.“ Auch KLINGLER (1982, 35) ist der Ansicht, dass Aufgabenstellungen klar und erfüllbar sein müssen, da unübersichtliche Anforderungen und allgemeine Anweisungen nicht zielführend sind und missverstanden werden können.

Schriftliche Aufgabenstellungen unterliegen einer „sprachlichen Wohlgeformtheit“ (Orthographie, Ausdruck, Grammatik), wobei bei Abweichungen von dieser der Schüler erwägen muss, ob dies zur Prüfungsanforderung gehört oder nicht (MEYERHÖFER 2003, 175f). Dies ist dem Lehrer vielleicht bei der Konstruktion seiner Aufgaben gar nicht bewusst und er erkennt nicht, dass Fehler aufgrund der Aufgabenstellung bei der Beantwortung der Fragen auftreten können.

Weiters ist zu beachten dass, bei einer größeren Aufgabenanzahl der Messfehler kleiner ist, und umgekehrt. Jedoch nähert man sich mit wachsender Aufgabenzahl der Ermüdungsgrenze und misst dann nicht mehr ausschließlich Fachleistung sondern auch allgemeine Belastbarkeit (SACHER 2004, 122f).

Die verschiedenen Aufgabenformen müssen im Unterricht geübt werden, da sonst die Bewältigung der ungewohnten Aufgabenform dem Schüler Schwierigkeiten bereiten kann. Dies hat wiederum Einfluss auf die Validität der Ergebnisse (SACHER 2004, 68).

Im Folgenden wird ein Einblick in die unterschiedlichen Aufgabentypen vorgenommen, um diese voneinander unterscheiden zu können.

II.3.2 Offene Aufgaben

Bei offenen Aufgaben beantwortet der Schüler die Frage nach freiem Ermessen (LIENERT, RAATZ 1998, 20). Offene Aufgabenformen erfordern frei formulierte Antworten mit denen das Verständnis geprüft werden kann (KLAUER 2001, 105), wobei der Umfang der geforderten Antworten stark variieren kann, von einem Satz bis hin zu einem ganzen Aufsatz (SACHER 2004, 62). Der Vorteil dieser Aufgabenform liegt nach KLAUER (2001, 105) darin, dass der Lernende selbst etwas formuliert, der Nachteil liegt in einer geringen Auswertungsobjektivität. Diese äußert sich darin, dass der Untersucher beziehungsweise der Lehrer über die Richtigkeit der Antwort entscheidet (LIENERT, RAATZ 1998, 8), dies subjektiv ist und von der persönlichen Meinung abhängt. Weiters trifft der Lehrer ein Urteil aufgrund seiner Fähigkeiten und seines Wertesystems (SCANNELL, TRACY 1977, 78).

In dieser Arbeit stehen offene Aufgaben und Multiple Choice Aufgaben im Zentrum der Betrachtung, trotzdem werden der Vollständigkeit halber auch halb offene Aufgaben vorgestellt.

II.3.3 Halb offene Aufgaben

Halb offene Aufgaben werden unterteilt in halb offene Freiantwortaufgaben, halb offene Ergänzungsaufgaben und halb offene Substitutionsaufgaben. Bei der Form der Freiantwortaufgaben sind kurze Antworten, in einer festgelegten Weise, zu beantworten. Bei den Ergänzungsaufgaben werden Informationen vervollständigt und bei den Substitutionsaufgaben vorgegebene Informationen verbessert oder ersetzt (SACHER 2004, 63f).

Offene und halb offene Formen dienen Lehrern dazu, Lösungswege nachvollziehen zu können (SACHER 2004, 66), was bei geschlossenen Aufgaben, die anschließend vorgestellt werden, kaum möglich ist.

II.3.4 Geschlossene Aufgaben

Zu geschlossenen Aufgaben können unter anderem Items (Aufgaben) mit vorgegebener Alternativantwort, Items mit mehreren Auswahlantworten, Umordnungsaufgaben, Zuordnungsaufgaben, sowie Ergänzungsaufgaben gezählt werden (GOETHALS 1994, 12f). Geschlossene Aufgaben haben, wie bereits erwähnt, verschiedene Antwortkategorien, die weiters unterteilt werden können.

Bei der Alternativvorgabe muss der Befragte sich zwischen zwei Antwortalternativen entscheiden. Im Gegensatz dazu muss er bei der Wahlvorgabe von mehreren vorgegebenen Antworten eine wählen. Zusätzlich gibt es noch Ratingskalen, bei denen die Antworten abgestuft werden, wie zum Beispiel sehr gut, gut und schlecht, sowie Rangordnungsaufgaben, wo die Antworten nach Präferenz sortiert werden (MITTELSTÄDT 2006, 78). In der Literatur findet man neben dem Begriff geschlossene Aufgaben auch den Begriff gebundene Aufgaben; gemeint ist dasselbe.

„Geschlossenen Aufgabenformen wird von Schulpraktikern gerne nachgesagt, damit lasse sich nur Wissen auf sehr einfachem Niveau überprüfen, nicht aber Verstehen.“ (SACHER 2004, 65) Auch mit Kurzantworten können laut Literatur eher Fakten erfragt werden als das Verständnis (SCHINDLER 1990, 29). Jedoch finden sich auch andere Statements in der Literatur, die den beiden vorangegangenen Argumenten widersprechen. Wie zum Beispiel: „Gut konstruierte Tests – ganz gleich, ob sie sich zusammensetzen aus Mehrfachwahlitems oder Freie-Antwort-Items – verlangen sowohl das Wiedererkennen von Informationen als auch ihre Anwendung auf neue Probleme.“ (SCANNELL, TRACY 1977, 78) Diese kontroversen Ergebnisse zwischen den verschiedenen Autoren, je nach dem welche

Aufgabenform sie bevorzugen, werden im Kapitel „Stand der Forschung“ noch näher diskutiert.

1970 wiesen bereits KRUMM und SEIDEL (zit. nach INGENKAMP 1989, 181) darauf hin, dass Klassenarbeiten und Prüfungsaufgaben in Berufsschulen hauptsächlich aus Wissensfragen und kaum aus Verständnisaufgaben bestehen. Auch anderen Orts findet man Hinweise darauf, dass sowohl offene Fragen als auch Multiple Choice Fragen eher Faktenwissen verlangen (FRIEDRICH 1987, 132).

Zum Erfassen kreativer Leistungen sind geschlossene Aufgaben nicht anwendbar, da die Antworten vorgegeben sind. Hierfür bieten sich offene Aufgabenformen an. Der Vorzug geschlossener Aufgaben liegt darin, dass mehr Aufgaben in der gleichen Zeit bearbeitet werden können, sodass Messfehler reduziert werden und die Reliabilität leichter zu sichern ist (SACHER 2004, 66f). Weiters ist „die Anonymität bei geschlossenen Frageformen [...] größer als bei offenen.“ (MITTELSTÄDT 2006, 78)

In der empirischen Sozialforschung werden geschlossene Aufgaben dann verwendet, wenn ein Vorwissen über den Informationsstand des Befragten besteht und die mögliche Meinung darüber Ziel der Untersuchung ist (FRIEDRICHS 1980, 199). Dabei kann es manchem Befragten nicht reichen, aus den gebotenen Antwortalternativen zu wählen, da er andere Gründe als die vorgegebenen für wichtig empfindet. Andere sehen vorgegebene Antworten wiederum als Erleichterung an, da sie sich nicht selbst Gründe einfallen lassen müssen (FRIEDRICHS 1980, 206).

II.3.5 Zusammenfassung: Aufgabenformen

„Zieht man eine [...] Summe aus dem Vergleich zwischen freier und gebundener Aufgabenform, so lässt sich keine Überlegenheit der einen oder der anderen Klasse von Aufgabenformen feststellen.“ (TRAUB, RURY 1990; zit. nach KLAUER 2001, 106)

Es zeigte sich, dass je nachdem was erfragt werden soll, entweder offene oder geschlossene Aufgaben geeigneter sind. Erfahrungen im Bereich der Aufgabenwahl haben hervor gebracht, dass es für die Motivation von Schülern nicht notwendig ist, verschiedene Aufgabenformen abwechselnd in einem Test zu verwenden (SCANNELL, TRACY 1977, 71).

Eine Befragung von Psychologiestudenten deutete darauf hin, dass die Befragten Kurzantwortfragen vor Multiple Choice Fragen bevorzugten (NIELD, WINTRE 2002, 95).

In den nächsten zwei Kapiteln wird näher auf Prüfungen mit offenen Fragen und Multiple Choice Fragen eingegangen, da sie für die vorliegende empirische Untersuchung eine wichtige Grundlage darstellen.

II.4 Offene Fragen

Offene Aufgabenformen werden von SACHER (2004, 62) in Freiantwortaufgaben, kurze Antworten, Essayform und umfangreichere Antworten gegliedert. Eine weitere Unterteilung findet sich auch bei SCANNELL und TRACY (1977, 81), die zwischen Kurz-Antwort-, Ergänzungs- und Essay-Test differenzieren. Kurzantwort- und Essay-Tests unterscheiden sich voneinander durch die Länge der geforderten Antwort. Bei der ersten Form wird ein Wort, ein Ausdruck oder höchstens zwei Sätze als Antwort gefordert, bei der zweiten Form dagegen mehrere Sätze. Beim Kurzaufsatz, auch Essaytest genannt, wird die Antwort durch die vorgegebenen Fragen und den eingeschränkten Platz eingegrenzt. Der Schwierigkeitsgrad solcher Fragen wird durch die Wahl der Formulierungen bestimmt. Zur Auswertung von Kurzaufsätzen müssen Kriterien vorgegeben werden, um die Objektivität zu erhöhen. Zum Beispiel könnte eine Frage fünf konkrete Antworten erfordern, wobei je nach Anzahl der Antworten die Punkte vergeben werden (ROLLETT 1997, 181f). Bei den Essay-Tests handelt es sich um die am weitesten verbreitete Form von Schultests, bei denen der Schüler sich an den Stoff erinnern und dann klare verständliche Antworten formulieren muss, um zu zeigen, dass er den Stoff verstanden hat (SCANNELL, TRACY 1977, 88).

II.4.1 Vorteile von offenen Fragen

Offene Aufgaben sind gekennzeichnet durch freie, nach eigenem Ermessen gegebene Antworten. Bei dieser Form wird dem Befragten die Möglichkeit gegeben, sein Wissen und Können frei anzuwenden und sich innerhalb der Prüfungszeit nach seinem Tempo, seinen Schwächen und Stärken zu richten (EBBINGHAUS, SCHMIDT 1999, 70ff).

In der empirischen Sozialforschung werden offene Fragen herangezogen, um den Bezugsrahmen des Befragten zu ermitteln, Informationen über die Einstellung von Personen zu gewinnen, sowie differenzierte Einstellungen zu erheben und damit Hypothesen über Zusammenhänge zu entwickeln (FRIEDRICHS 1980, 199). Hier sei angemerkt, dass die Reihenfolge, mit der jemand eine offene Frage beantwortet, nicht immer etwas über die Wichtigkeit des Niedergeschriebenen aussagt (FRIEDRICHS 1980, 206).

Offene Fragen erschweren auf der einen Seite die Auswertung und beeinträchtigen die Auswertungsobjektivität, auf der anderen Seite werden dadurch kreative Lösungen möglich (GOETHALS 1994, 13).

BRIDGEMANN (1992, 269) ist der Ansicht, "if the intent to the test is to describe specific skills that students possess, the open-ended format seems to be clearly superior." Er begründet dies damit, dass nicht alle falschen Antworten in einem Multiple Choice Test enthalten sein können, die bei offenen Antworten auftreten könnten. Er ist der Meinung, dass die Befragten bei offenen Fragen die Antworten kreieren müssen und nicht nur eine von zum Beispiel fünf vorgegebenen wählen müssen (BRIDGEMANN 1992, 271).

II.4.2 Nachteile von offenen Fragen

Der Nachteil von solchen Fragen liegt darin, dass Personen, die sich schriftlich besser ausdrücken, bevorzugt werden (EBBINGHAUS, SCHMIDT 1999, 70ff). In Bezug auf die Auswertungsobjektivität von Aufsätzen konnten in Untersuchungen in den USA und England Hinweise darauf gefunden werden, dass Aufsätze von verschiedenen Beurteilern sehr unterschiedlich bewertet werden, dass Lehrer grammatikalische und orthographische Fehler unterschiedlichst in die Bewertung mit einbeziehen und die gleichen Beurteiler die gleichen Aufsätze zu verschiedenen Zeitpunkten ebenso unterschiedlich bewerten (INGENKAMP 1989, 182).

Bei diesem Fragentyp muss nicht auf Formulierungen geachtet werden, in dem Sinne, dass sie nicht wiedererkannt werden, sondern frei wiedergegeben werden können. Die Gefahr besteht darin, dass der Schüler nur sehr oberflächlich lernt, weil er dem Lehrer etwas vortäuschen kann (SCHINDLER 1990, 26f).

II.4.3 Zusammenfassung: Offene Fragen

Die Konstruktion der Aufgaben im Multiple Choice Format ist aufwendiger, als die von offenen Fragen, jedoch ist die Auswertung von offenen Fragen aufwendig und von geringer Objektivität (BESSOTH 2002, 33f). Die Entscheidung für die Form der Fragen, je nachdem welche Ergebnisse – eigenständige Lösungen oder Einstellungen – er haben möchte, obliegt dem Prüfer. Dieser wägt die Vor- und Nachteile der Formate für seine Fragestellung ab.

II.5 Multiple Choice Fragen

In dieser Arbeit werden unter Multiple Choice Aufgaben jene verstanden, bei denen Items mit mehreren Auswahlantworten vorgegeben werden. SACHER (2004, 64) bezeichnet diese Form der Aufgaben als geschlossene Identifikations-Auswahlaufgaben, genauer gesagt, ist die richtige, beziehungsweise die falsche Lösung auszuwählen oder zu identifizieren. Um noch konkreter zu differenzieren muss gesagt werden, dass es sich bei Multiple Choice Aufgaben auch um so genannte „Mehrfach-Wahl-Aufgaben“ handeln kann, „bei denen nicht eine, sondern mehrere Lösungen gewählt werden müssen.“ (INGENKAMP 1971, 51)

Ein Multiple Choice Item besteht immer aus einem Stamm, der Prüfungsfrage und den Alternativantworten, anhand derer der Schüler seine Leistung durch ankreuzen demonstrieren kann (GAGE, BERLINER 1996, 647). Neben den richtigen Antworten werden auch falsche Antworten vorgegeben, so genannte Distraktoren. Distraktoren dienen dazu, dass Schüler, die die Antwort nicht wissen, die falsche Antwortalternative mit etwa der gleichen Wahrscheinlichkeit auswählen, wie die Richtige (LIENERT, RAATZ 1998, 26). Das heißt, Schüler die bei einem Multiple Choice Test gute Leistungen erzielen, sollten in der Lage sein, „die ablenkenden Alternativen aufgrund ihrer Kenntnisse der Lernziele des Unterrichts auszuschließen“ (SCANNELL, TRACY 1977, 129). Bei der richtigen Wahl der Distraktoren halten uninformierte Personen alle Antwortalternativen für gleich wahrscheinlich. Suboptimales Antwortverhalten entsteht durch ziemlich offensichtlich falsche Antworten. Das heißt, dass der Geprüfte die falschen Antworten eliminieren kann aufgrund ihres falschen Inhaltes (GOETHALS 1994, 11). Eine neutrale Option als Antwortmöglichkeit verhindert das Treffen einer Entscheidung und die Chance, „durch Raten auf der Grundlage von partiellen Wissen Punkte zu sammeln“ (KLAUER 2001, 109). Durch Ausschalten der Distraktoren zur richtigen Antwort zu gelangen, erweist sich teilweise schwieriger, als durch Gedächtnisassoziationen beziehungsweise durch Teilwissen diese zu finden (INGENKAMP 1989, 187).

Bei diesem Antwortformat sollte der Lernaufwand nicht unterschätzt werden, da bluffen nicht sinnvoll ist und man eine klare Entscheidung treffen muss (SCHINDLER 1990, 24).

II.5.1 Multiple Choice Tests

„Unter informellen Tests werden meist jene Formen schriftlicher Überprüfungen verstanden, die vom Lehrer selbst erstellt werden, sich an den bearbeitenden Lehrzielen orientieren und Aufgaben zu diesen Lehrzielen enthalten.“ (SCHEIBER 1990, 103) Formelle oder auch Formative Tests erfüllen einen Diagnosezweck und decken mehr Lehr- und Lernstoff als informelle Tests ab (SCHWENDENWEIN 2000, 178).

In dieser Arbeit werden unter Multiple Choice Tests informelle Tests verstanden, bei denen eine Frage und mehrere Antworten vorgegeben werden, wobei eine, beziehungsweise mehrere Antworten richtig sein können. Im Folgenden wird auf die Vor- und Nachteile dieser Prüfungsform eingegangen und der Hauptkritikpunkt, die Ratewahrscheinlichkeit, näher erörtert.

II.5.2 Vorteile von Multiple Choice Fragen

Neben der Überprüfung angelernten Wissens können komplexere Zusammenhänge, sowie Erfolgskontrollen im affektiv-sozialen Bereich, durch abgestufte Antwortmöglichkeiten mit Multiple Choice Tests erfragt werden (HEPP, BENZ 1987, 31). RAPP (1971, 74ff) ist der Ansicht, dass es von der Aufgabenstellung abhängt, ob bei Schülern höhere geistige Prozesse aktiviert werden, wie das Verstehen der Terminologie, das Verstehen von Gesetzmäßigkeiten, das Verstehen von Beziehungen und induktives Denken.

Ein Test ist ökonomisch, wenn er schnell durchzuführen ist, als Gruppentest verwendet werden kann und schnell auszuwerten ist (LIENERT, RAATZ 1998, 12). Neben einem hohen Maß an Objektivität sind Multiple Choice Tests auch arbeitsökonomisch. Durch diese Art der Tests können nicht nur Faktenwissen, sondern auch komplexere Problemstrukturen erfasst werden, also im Spezielleren kognitive wie auch affektiv-soziale Lernziele überprüft werden (HEPP, BENZ 1987, 29). Kognitive Lernziele werden unter anderem in die Kategorien Wissen, Denken und Beurteilen unterteilt. Unter affektiven Lernzielen werden unter anderem Gefühle, Emotionen und Wertschätzungen verstanden (GAGE, BERLINER 1986, 68f). Multiple Choice Fragen wurden entwickelt, um bei einer Prüfung Ergebnisse zu erzielen, die vom Prüfer möglichst unabhängig sind (INGENKAMP 1989, 181). Das heißt, egal wer die Prüfung durchführt oder auswertet, die Ergebnisse sind immer dieselben. Aufgrund der vergleichbaren Ergebnisse bei Multiple Choice Tests, erhöht sich die Auswertungsobjektivität (SACHER 2004, 67). „In Bereichen der Leistungsdiagnostik, wo, (hauptsächlich) aus Gründen der Testdurchführungs- oder Auswertungsökonomie, Gruppen- oder Computertests eingesetzt werden, kommt kaum ein Autor an der Multiple-choice-Vorgabe vorbei.“ (GOETHALS 1994, 90)

Neben der Einfachheit der Durchführung und der Tatsache, dass mehrere Schüler gleichzeitig geprüft werden können, besteht bei diesem Format die Möglichkeit, aufgrund einer automatisierten Auswertung (per Computer oder Schablone), Zeit und Kosten zu sparen (DUFRESNE, u.a. 2002, 174).

Die Vorteile von Multiple Choice Tests machten diese in der Psychologie und Pädagogik zu wichtigen Messinstrumenten. Vor allem in den USA werden unter anderem Abschlusstests in der High School und Aufnahmeprüfungen für die Universität in Form von Multiple Choice Tests vorgegeben (BAR-HILLEL, u.a. 2005, 4).

II.5.3 Nachteile von Multiple Choice Fragen

Ein Nachteil von Multiple Choice Tests ist die mitunter eingeschränkte Reliabilität. VALE (1978; zit. nach GOETHALS 1994, 14) konnte Hinweise darauf finden, dass bei der freien Beantwortung Fehler durch Raten vermieden werden können. Nach BUSE (1977, 546) ist ein Test um so reliabler, je weniger Personen raten und je geringer Zufallstreffer möglich sind. Die Ratemöglichkeit kann durch eine größere Anzahl von Items und eine höhere Anzahl von Distraktoren verringert werden, was jedoch gleichzeitig Einbußen in Bezug auf die Testökonomie mit sich bringt (GOETHALS 1994, 16).

Bei Multiple Choice Aufgaben sind die Schüler, die raten, von denjenigen, die die Angabe falsch verstehen, nicht zu unterscheiden (MEYERHÖFER 2003, 16). Raten und Wissen können bei diesen Aufgaben nicht voneinander getrennt werden und was von beiden zur Lösung führt kann nicht festgestellt werden (MEYERHÖFER 2003, 233). Multiple Choice Fragen ermöglichen in manchen Fällen nach MEYERHÖFER (2003, 193) das Lösen dieser ohne jedes inhaltliche Verstehen. Durch Raten kann mittels partiellen Wissens die eine oder andere Antwortmöglichkeit als unwahrscheinlich ausgeschlossen werden. Für die Lernenden, alle haben die Möglichkeit zu raten, entstehen keine Vor- oder Nachteile. Prüfer versuchen die Ratewahrscheinlichkeit zu verringern, in dem sie mehrere Testaufgaben geben (KLAUER 2001, 106).

Die Grenzen in der Anwendung von Multiple Choice Tests sind erstens abhängig von der Erfahrung des Konstrukteurs und zweitens kann mit dieser Form der Aufgaben sprachliche Spontanleistung nicht gezeigt und Lösungswege kaum erfasst werden (INGENKAMP 1989, 189/203). KLAUER (2001, 109) ist jedoch der Ansicht, dass mittels Multiple Choice Aufgaben neben der Überprüfung, ob ein Schüler sein Wissen abrufen kann, auch Problemlösungen erfasst werden.

„Die Kritik, Antwort-Auswahl-Aufgaben seien pädagogisch nicht zu verantworten, weil sie auch falsche Antworten vorlegten, stammt von einer didaktisch völlig überholten Position und steht im Gegensatz zu dem Bestreben, Schüler zur kritischen Prüfung und Selbstüberprüfung anzuregen.“ (INGENKAMP 1989, 190)

Empirisch konnte nicht bestätigt werden, dass Schüler aufgrund falscher Antworten falsche Informationen lernen (PRESTON 1965; zit. nach LIENERT, RAATZ 1998, 26).

II.5.4 Ratewahrscheinlichkeit

EDEL (1965; zit. nach INGENKAMP 1989, 187) schrieb: „Ein objektiver Test ermöglicht und fördert gelegentlich das Raten. Eine Aufsatzprüfung ermöglicht und fördert gelegentlich das Bluffen.“

Bei der Konstruktion von Multiple Choice Fragen sollte laut ROLLETT (1997, 185f) darauf geachtet werden, dass die richtigen Antworten stets in einer anderen Position stehen, die Distraktoren plausibel erscheinen und sich nicht in ihrer Formulierung von den richtigen Antworten unterscheiden. Ebenso sollten ihrer Meinung nach doppelte Verneinungen vermieden werden, da in diesem Fall nicht der Inhalt, sondern die grammatische Kompetenz der Schüler abgeprüft wird und somit die Validität der Prüfungsfrage eingeschränkt wird. Auch HEPP und BENZ (1987, 31) fordern eindeutige Aussagen, plausible Antworten, die im Inhalt nicht unmöglich erscheinen. “[...] the ability to answer the question should not depend on general knowledge (i.e. information going beyond what is given in the text itself).” (PERRIN 2004, 173)

„Für alle Auswahlantwortformen gilt, daß die Ratewahrscheinlichkeit umso geringer wird, je mehr plausible, aber falsche Antwortalternativen (sogenannte „Distraktoren“) vorgegeben werden“. (ROLLETT 1997, 184) Bei fünf möglichen Antworten liegt die Ratewahrscheinlichkeit bei 20% (INGENKAMP 1989, 186). „Der mögliche Nachteil eines blinden Ratens bei Antwort-Auswahl-Aufgaben wird durch den Vorteil einer viel größeren Stichprobe möglicher Fragen mehr als ausgeglichen.“ (INGENKAMP 1989, 188) Als Beispiel hierfür sei angemerkt, dass, wenn ein Schüler 50% des gelernten Stoffgebietes beherrscht, ihm aber nur 10 Fragen vorgelegt werden, die Wahrscheinlichkeit hoch ist, dass er die Fragen nicht weiß und errät. Sind dagegen mehr Fragen vorgegeben, besteht für den

Schüler die Chance mehr zu wissen (INGENKAMP 1989, 188). Außerdem dienen die Distraktoren dazu, die Antworten genau zu prüfen und diese zu hinterfragen (FRIEDRICH 1987, 142).

Gerade bei Multiple Choice Tests wird immer wieder die Problematik des Ratens als Nachteil dieses Verfahrens aufgegriffen, da zwischen Raten und Wissen kaum unterschieden werden kann. Um den Anreiz des Ratens zu minimieren, wird von vielen Testern bei der Bewertung dieser Testform folgendes Punktesystem verwendet: Bei der richtigen Antwort wird ein Punkt, bei einer falschen Antwort einer abgezogen und bei keiner Antwort werden null Punkte vergeben (BAR-HILLEL, u.a. 2005, 5). „Zwischen "Raten" und "Wissen" gibt es fließende Übergänge (informiertes Raten, Teilwissen).“ (FRIEDRICH 1987, 137)

II.5.5 Zusammenfassung: Multiple Choice Fragen

Es konnte gezeigt werden, dass Multiple Choice Tests in ihrer Durchführung und Auswertung weitestgehend objektiv sind, da es sich um vorgegebene Antworten handelt. Jedoch wurde festgestellt, dass die Wahl der Antworten nicht sehr einfach ist, da zu einer richtigen Antwort adäquate Distraktoren gefunden werden müssen. Bei schlecht gewählten Distraktoren erhöht sich die Ratewahrscheinlichkeit, was zu Verfälschungen des Testergebnisses führen kann. Wie und was bei diesen Tests geprüft werden kann, ob es sich um reines Wiedererkennen handelt oder höhere Denkprozesse angeregt werden, das Gelernte von Dauer ist oder schnell vergessen wird, wird in der Literatur sehr unterschiedlich dargestellt. Im nächsten Kapitel werden zum Thema Multiple Choice Tests diverse Untersuchungen näher vorgestellt.

II.6 Stand der Forschung

In diesem Kapitel werden die unterschiedlichsten Untersuchungen im Zusammenhang mit Multiple Choice Fragen und Tests vorgestellt. Wie sich zeigt gibt es kaum Untersuchungen zu diesem Thema im deutschsprachigen Raum, was darauf schließen lässt, dass es sich hierbei um ein zum Teil unerforschtes Gebiet handelt.

II.6.1 Multiple Choice versus Offene Fragen

Bei einem Vergleich zwischen einem Multiple Choice Test und Kurzantwort Test, welche beide von Lehrern konstruiert wurden, ging es darum zu überprüfen, inwiefern sie "retention learning" unterstützen. Unter "retention learning" wurde lernen verstanden, "which lasts beyond the initial testing and it is assessed with tests administered 2 or more weeks after the information has been taught and tested." (HAYNIE 1994, 33) Die Schüler wurden per Zufall in drei Gruppen eingeteilt, wobei die erste Gruppe mittels Multiple Choice Test und die zweite mittels Kurzantworten getestet wurde. Die dritte stellte die Kontrollgruppe dar, die keinen Test hatte, mit der Begründung es sei noch keiner entwickelt worden. Alle Probanden bekamen eine Lernvorlage, mit der Anweisung diese wie jedes andere Skript zu lernen. Der erste Test der durchgeführt wurde bestand aus 20 Items, die bei der ersten Gruppe im Multiple Choice Format und bei der zweiten im Kurzantwortformat vorgegeben wurde. Die Items waren bei beiden Gruppen gleich. Ein verspäteter Test, der frühere und neue Informationen enthielt bestand aus 30 Items im Multiple Choice Format, wobei 20 dieser Items vom ersten Test und 10 neu waren. Dieser Test wurde allen drei Gruppen vorgelegt. Es konnten Hinweise gefunden werden, dass der "initial test" im Multiple Choice Format "retention learning" mehr fördert als mit Kurzantworten. Dies könnte auf der einen Seite daran liegen, dass der erste und zweite Test im Multiple Choice Format war. Auf der anderen Seite könnte dies auch zustande gekommen sein, weil die

Antworten vorgegeben waren und bei den Kurzantworten kein Hinweis auf die Antworten gegeben war. Keine Unterschiede zwischen den beiden Experimentalgruppen zeigten sich in Bezug auf die 10 neuen Items im zweiten Test, wobei beide Gruppen besser abschnitten als die Kontrollgruppe. (HAYNIE 1994, 32ff) “[...] short-answer and multiple-choice tests were both equally effective in promotion of retention of information which was studied but which was not actually reflected in the test items.” (HAYNIE 1994, 41)

SAX und LEVERNE (1968, 170ff) erforschten, ob “individuals trained on multiple-choice tests will obtain higher mean scores on both the multiple-choice and recall examinations”. Diese “recall examinations” sind Tests mit kurzen und Vollendungsantworten. Untersucht wurden Studenten von zwei Vorlesungen. Den Studenten, die in Vorlesung 1 gingen, wurde mitgeteilt, dass sie drei Tests und einen Abschlusstest mit kurzen Antworten bekämen. Denjenigen die Vorlesung 2 besuchten, wurde gesagt, dass es drei Multiple Choice Tests und ein Abschlusstest in eben diesem Verfahren sein werden. Die Fragen waren in beiden Vorlesungen gleich und unterschieden sich, wie bereits erwähnt, nur durch die Art des Antwortformats. Beide Vorlesungen wurden gleich unterrichtet, die Studenten beider Vorlesungen bekamen dieselben Instruktionen und die gleichen Materialien zur Verfügung gestellt. Der Abschlusstest setzte sich in beiden Gruppen sowohl aus “recall” als auch Multiple Choice Antworten zusammen, worüber die Studenten informiert wurden. Das heißt, dass alle Studenten beim Abschlusstest in zwei Gruppen (A, B) geteilt wurden. Gruppe A bekam einen “recall” Test und Gruppe B einen Multiple Choice Test. Die Ergebnisse wiesen darauf hin, dass aufgrund des Trainings mit Multiple Choice Items, die Leistungen beim Multiple Choice Abschlusstest höher waren, als beim Training mit “recall” Items. Weiters zeigte sich, dass es schwieriger war “to bluff on a closely related options requiring fine discriminations than it is to bluff on recall-type items”. (SAX , LEVERNE 1968, 172)

Eine andere Studie beschäftigte sich mit den Unterschieden im Antwortverhalten von Schülern bei Bruchrechnungen “on the diagnosis of examinee misconceptions in a procedural task” (BIRENBAUM, TATSUOKA 1987, 385ff). Die Rechnungen wurden im Multiple Choice Format und im offenen Fragen Format gestellt. Diese Erhebung deutete darauf hin, dass Schüler, die die Aufgabe nicht lösen konnten dazu neigten, im Multiple Choice Format weniger “their rules of operation for solving procedural tasks” anzuwenden als bei offenen Fragen (BIRENBAUM, TATSUOKA 1987, 392). Bei den offenen Fragen mussten die Schüler die ganze Rechnung rechnen, bei den Multiple Choice Fragen mussten sie die Distraktoren finden, wobei die falschen Antworten aus Fehlern bei zuvor gestellten offenen Fragen bestanden. Resümee dieser Studie war, dass Multiple Choice Fragen einfacher zu bewerten sind, jedoch nicht die adäquaten Informationen bringen, um “students’ misconceptions with respect to the given subject matter” aufzuzeigen (BIRENBAUM, TATSUOKA 1987, 392ff). Hierfür eignen sich den Autoren nach offene Fragen.

II.6.2 Einschätzung und Präferenz

Bei einer Befragung von Studenten zum Thema Multiple Choice Prüfungen, konnten die Autorinnen folgende Tendenzen herausfinden: Multiple Choice Prüfungen werden im Gegensatz zu offenen Fragen als schlechter eingestuft. Studenten gaben an, sich auf Multiple Choice Prüfungen primär mit Musterklausuren und mit Hilfe des „Wiedererkennfaktors“ vorzubereiten. Bei offenen Fragen wird auf Verständnis gelernt und mit Hilfe von Zusammenfassungen. Von 24 Befragten, gaben 6 an, bei Multiple Choice Tests mit kaum einer Vorbereitung anzutreten. Bei der Frage nach der Dauerhaftigkeit des Prüfungswissens konnten keine Unterschiede zwischen den beiden Formaten herausgefunden werden (GANNESHOFFER, KIFFEL 2005, 62ff).

Kann mittels Multiple Choice Tests mehr als nur Reproduktion erfasst werden? Wie sehen Schüler das und wie bereiten sie sich auf Prüfungen vor? Diesen Fragen gingen SCOLLER und PROSSER (1994, 268ff) nach. In der Literatur findet man zu diesen Fragen kontroverse Ergebnisse. Auf der einen Seite wurde dokumentiert, Studenten lernten unabhängig vom Prüfungsformat und auf der anderen Seite gab es Befunde dafür, dass Studenten intensiver für offene als für geschlossene Fragen lernten. "Research has generally associated deep learning approaches with higher achievement (both in terms of higher quality outcomes and grades) and greater learner satisfaction, and surface learning approaches with lower achievement (poor grades) and greater learner dissatisfaction." (SCOLLER, PROSSER 1994, 269) Im Zuge dieser Erhebung wurden insgesamt 190 Studenten befragt. 69 Studenten besuchten Kurs A, waren bereits im zweiten Studienjahr und hatten bereits Erfahrungen mit Prüfungen an der Universität. Bei diesem Kurs bestand der Abschlusstest teilweise aus Multiple Choice Fragen, die 20% der Note ausmachten. Die restlichen 121 Studenten waren im ersten Studienjahr, 34 von ihnen besuchten Kurs B, der ein Semester lang dauerte und dessen Abschlusstest ausschließlich aus Multiple Choice Fragen bestand. 87 Studenten besuchten Kurs C, der über zwei Semester abgehalten wurde. Die Studenten dieses Kurses konnten bereits im ersten Semester Erfahrungen mit Multiple Choice Tests sammeln. Diesen 190 Studenten wurde ein Fragebogen drei Wochen vor ihrer Abschlussprüfung, die aus Multiple Choice Fragen bestand, vorgelegt. Der Fragebogen beinhaltete Fragen dazu, wie die Studenten im Allgemeinen lernten, ihre Strategien speziell für Multiple Choice Tests und wie sie sich für die Abschlussprüfung vorbereiteten. Die Ergebnisse der Fragebögen wurden in einem ersten Schritt alleine betrachtet und in einem zweiten im Zusammenhang mit den Testergebnissen.

Die Ergebnisse dieser Untersuchung deuteten darauf hin, dass bessere Noten und gut geplante Lernstrategien verbunden waren mit "deep and/or achieving general orientations to study", was bei allen drei Kursen gleich war. Studenten die oberflächlich lernten, schienen keine Strategien zu haben, wie sie sich für den Test vorbereiten könnten und hatten "incompatible conceptions of the difference

between reproduction and understanding" (SCOULLER, PROSSER 1994, 277). Weiters konnte ein Zusammenhang zwischen der Wahrnehmung, dass mit Multiple Choice Tests höhere kognitive Funktionen, wie Anwendung und Analyse von Wissen verlangt würden und dem Vorsatz der Studierenden, sich gute Lernstrategien zu überlegen, gefunden werden (SCOULLER, PROSSER 1994, 277f).

Inwiefern Studenten ein Frageformat bevorzugen und dies mit ihrem Lernverhalten und mit ihrer Wahrnehmung "of the intellectual abilities being assessed" zusammenhängt wurde von SCOULLER (1998, 457) näher untersucht. Die Daten wurden mittels eines Fragebogens und den Noten erhoben. Der Fragebogen beinhaltete Fragen zu der Lernmethode, zu dem wahrgenommenen Level der "intellectual abilities" und der Präferenz für eine Erhebungsmethode. Diese Methoden waren in dem Kurs an dem die Probanden teilnahmen ein Multiple Choice Test und ein "assignment essay". Beide zusammen ergaben die Note. Die Ergebnisse dieser Untersuchung zeigten, dass Studenten oberflächlicher für Multiple Choice Tests lernten und diese Tests niedrigere "intellectual abilities" verlangten. Unter oberflächlicher wurde das Abrufen und Wiedererkennen von Informationen verstanden, bei dem Teilinformationen ausgebildet wurden. Im Gegensatz dazu standen höhere Lernprozesse, die auf Verstehen, Anwendung und Zusammenhänge bilden, beruhten. Ein Großteil der befragten Studenten bevorzugte die "assignment essay" und sie schnitten auch besser bei diesen ab. Jene die Multiple Choice Tests bevorzugten, waren der Ansicht, dass dieses Format höhere kognitive Prozesse anregte. Diese Studenten schnitten beim Multiple Choice Test besser ab (SCOULLER 1998, 453ff). "[...] the non employment of deep preparation strategies may be predictive of higher MCQ examination scores." (SCOULLER 1998, 467) Es konnten Hinweise gefunden werden, dass die Präferenz für ein Format einen Einfluss auf das Abschneiden bei diesem hatte. Für den Kurs den die Studenten bei dieser Untersuchung besuchten, zeigte sich, dass der "assignment essay" die bessere Prüfungsmethode darstellte, da die Studenten bei diesem Verfahren ihre Fähigkeiten demonstrieren konnten (SCOULLER 1998, 468f).

CHAN und KENNEDY (2002, 959ff) gingen der Frage nach, ob Multiple Choice Tests, im Gegensatz zu Tests mit "constructed response" (CR), von Wirtschaftsstudenten als einfacher wahrgenommen werden. Es wurde im Vorfeld überlegt, ob eventuell die Möglichkeit des Ratens oder die Tatsache, dass man sich aufgrund der vorgegebenen Antworten durchs Ausschlussprinzip an die Richtige herantasten kann, die Einfachheit erklären könnte. Im Rahmen ihrer Untersuchung führten sie eine Literaturreview durch, wobei Erhebungen gefunden wurden, bei denen Multiple Choice Tests als einfacher und jene bei denen konstruierte und Multiple Choice Antworten als gleich schwer dargestellt wurden. Unter "constructed response" wurden Antworten verstanden, bei denen ein Wort, eine Phrase oder ein Kurzaufsatz die Lösung darstellte. Es wurden 196 Studenten im Fach Wirtschaft geprüft. Wobei zwei Prüfungen, A und B, zur selben Zeit stattfanden. A und B beinhalteten jeweils 36 Multiple Choice, 12 "special CR" und 7 andere CR Fragen. Bei den "special CR" wurden die Studenten aufgefordert, eine Zeile als Antwort zu geben. Acht Fragen waren in beiden Gruppen gleich. Eine weitere Differenzierung wurde von den Wissenschaftlern vorgenommen: Die Multiple Choice Fragen von Prüfung A waren bei Prüfung B die "special CR" Fragen und "special CR" von A waren die Multiple Choice Fragen von B. Mit Hilfe dieses Designs wurden die Daten erhoben, wobei folgende Ergebnisse gefunden werden konnten. Bei den acht gleichen Fragen in beiden Formaten konnten keine Unterschiede zwischen guten und schlechten Schülern festgestellt werden. Bei den restlichen Fragen zeigte sich, dass gute und schlechte Schüler besser bei Multiple Choice Fragen abschnitten. Insgesamt schnitten die Studenten nach der Ratekorrektur besser bei Multiple Choice Fragen ab, als bei denselben Fragen im CR Format. Weiters konnten Hinweise gefunden werden, dass bei einer guten Wahl der Distraktoren, diese Antwortform Schwierigkeiten bereitete, "by causing students to worry about erroneous factors that they otherwise would not have taken into consideration". (CHAN, KENNEDY 2002, 970)

Der Zweck der Studie von ROWLEY (1974, 15ff) lag darin zu beschreiben "whether certain selected student characteristics are associated with success on

multiple choice tests to an extent which cannot be accounted for by their association with success on a free response tests covering identical subject matter." Die für diese Untersuchung gewählten Eigenschaften waren Leistungsmotivation, Testangst, Risikoverhalten und "testwiseness". Die aufgestellte Hypothese lautete: Im Vergleich zu offenen Fragen, werden von Studierenden, die sehr motiviert sind, keine Angst haben, "testwise" und risikobereit sind, Multiple Choice Fragen bevorzugt. 154 Schüler wurden an zwei Tagen im Abstand von fünf Wochen getestet. Am ersten Tag bekamen sie offene Fragen zum Stoff zusammen mit einem Fragebogen mit 30 Fragen aus der "Russell Achievement Motivation Scale" (1969) und 19 Fragen vom "Achievement Anxiety Test" von ALPERT und HABER (1960). Die Befragten wurden nicht darüber informiert, dass ein weiterer Test folgen würde. Fünf Wochen später wurde den Schülern gesagt, "that the multiple choice test was being given because it was thought that they might have been able to do better on that type of test than on the one previously administered." (ROWLEY 1974, 18)

Der Multiple Choice Test setzte sich aus 30 Vokabelfragen, 16 Mathematikfragen, 16 Fragen aus der "Slakter's testwiseness scale" (1970) und 18 aus den Bereichen Geschichte, Geographie, Allgemeinwissen, Naturwissenschaft und Mathematik zusammen. Analysiert wurden die Zusammenhänge zwischen den Eigenschaften der Studenten und den Multiple Choice Ergebnissen. Es konnten keine Belege gefunden werden, dass leistungsmotivierte oder ängstliche Schüler Multiple Choice Tests bevorzugen oder ablehnen. Die Vokabelfragen deuteten darauf hin, dass jene Schüler bei diesem Format einen Vorteil hatten, die "high on testwiseness and/or risktaking" waren. Kein Unterschied zwischen Multiple Choice und offenen Fragen konnte in Bezug auf die Mathematikitems gefunden werden. Die Differenzen in den Ergebnissen wiesen darauf hin, dass die Wahl der Antwortformate vom Gegenstand, der untersucht werden soll, abhing (ROWLEY 1974, 18ff).

Andere Untersuchungen zeigen, dass Studenten hilflos und frustriert sind, wenn es darum geht, eine Antwort, und zwar die beste, anzukreuzen (NIELD, WINTRE 2002, 95).

II.6.3 Speziell zum Thema Multiple Choice Tests

NIELD und WINTRE (2002, 96ff) gingen der Frage nach, wie Studenten Multiple Choice Fragen kommentieren? Dazu gaben sie 201 Studenten Multiple Choice Tests inklusive eines Antwortbogens vor. Auf diesem mussten die Studenten die für sie „beste“ Antwort ankreuzen und hatten die Möglichkeit ihre Antworten zu erläutern, wofür sie Plus- oder Minuspunkte bekamen. Weiters wurde den Studenten ein anonymer Fragebogen vorgelegt, mit dessen Hilfe sie nach ihrer Einstellung zu der Erläuterungsoption und ihrer Einstellung zu Multiple Choice Tests, befragt wurden. Es zeigte sich, dass die Befragten Multiple Choice Fragen und Richtig-/Falsch Antworten favorisierten, da sie bei diesen gegebenenfalls raten konnten und ihrer Meinung nach wenig geistige und körperliche Anstrengung erforderten. 70% der Befragten gaben an, dass die Erläuterungsoption Vorteile hatte, da diese ihre Angst und Frustration relativierte. Jedoch wurden auch Aufsätze und Kurzantworten von manchen Studenten bevorzugt, da sie mit diesen zeigen konnten, was sie gelernt hatten (NIELD, WINTRE 2002, 96ff).

ATTALI und BAR-HILLEL (2003, 109) erörtern das Problem der Positionierung von richtigen Antworten bei Multiple Choice Tests. Sie fanden in der Literatur eine Studie von BERG und RAPAPORT (1954; zit. nach ATTALI, BAR-HILLEL 2003, 111) bei der Studenten in einem Kurs aufgefordert wurden, wöchentlich eine Multiple Choice Frage zu konstruieren. Es zeigte sich, dass in 80% der entwickelten Fragen (von 1000), die richtige Antwort in der Mitte stand. Weiters untersuchten BERG und RAPAPORT (1954; zit. nach ATTALI, BAR-HILLEL 2003, 111) das Antwortverhalten von Studenten bei inhaltslosen Fragen und Antworten. Dazu legten sie 374 Studenten zum Beispiel Frage 2 mit Antwort A, B, C, D, also Fragen und Antworten ohne konkreten Inhalt nur mit Zahlen und Buchstaben, und 203 Studenten inhaltslose Fragen vor, bei denen die Antworten in umgekehrter Reihenfolge vorgegeben wurden, D, C, B, A. In 70% der Fälle wurde die richtige Antwort in der Mitte angekreuzt.

Daraufhin baten ATTALI und BAR-HILLEL (2003, 111ff) 190 Personen eine Multiple Choice Frage aufzuschreiben und vier Antwortmöglichkeiten zu formulieren. 65 von ihnen hatten Erfahrung in der Konstruktion von Multiple Choice Fragen, fast 80% der richtigen Antworten lagen in der Mitte. Bei den anderen Probanden, die keine Vorkenntnisse hatten, fast 70%. Die Befragten gaben an, dass sie sich kaum Gedanken über die Positionierung der richtigen Antwort machten. Die Fragen wurden nicht anhand der Position, sondern aufgrund des Inhalts beantwortet. Da die Position beim Raten eine Rolle spielen könnte, gaben sie 127 Studenten zwei Fragen und jeweils vier Antwortoptionen (A, B, C, D) ohne Inhalt vor und wollten anhand dieses Designs von den Studenten wissen, wo die richtige Antwort stehen könnte. Wieder 80% der Befragten tendierten bei der richtigen Antwort zur Mitte. Bei dieser Untersuchung konnte beobachtet werden, dass "[...] a tendency toward the middle—or away from the edges—exists in contexts that have nothing to do with tests or with experience." (ATTALI, BAR-HILLEL 2003, 114)

BALCH (2002, 106ff) betrachtete dagegen näher die Reihenfolge von Fragen bei Multiple Choice Tests. Er gab dieselben Fragen mit denselben Antworten, beide im Zuge von drei gleichen Tests unterschiedlichst positioniert, Studenten vor. Beim ersten Test wurden die Fragen in der Reihenfolge der Vorlesungsinhalte (S), beim zweiten Test wurden Kapitel und innerhalb der Kapitel die Fragen per Zufall (CC) vorgegeben und beim dritten Test wurden sie vom Computer per Zufall (R) geordnet. Es zeigte sich, dass die Ergebnisse der Tests durch die Reihenfolge der Items beeinflusst wurden. Die Studenten erzielten etwas bessere Resultate bei der Vorgabe nach den Vorlesungsinhalten, als bei der nach Zufall, was in der Praxis dazu führen könnte, dass der Prüfer diese Form der "item order" verwendet, um die best möglichen Leistungen seiner Studenten zu erzielen (BALCH 2002, 106ff).

NEELY und andere (2002, 108ff) gingen ebenfalls der Frage nach, ob die Item Reihenfolge die Testleistung in einem Multiple Choice Test beeinflussen würde. Sie vermuteten, dass die Positionierung der Fragen aufgrund der Vorlesungsinhalte ängstliche Studenten fördern würde, da diese Art der Aufgabenzusammenstellung eine bessere Struktur liefert und einen klareren Fokus auf die Aufgabe legen würde. Es konnte nicht nachgewiesen werden, dass

die Reihenfolge mit Ängstlichkeit zusammenhing, jedoch konnten folgende Tendenzen gefunden werden: "High-anxiety students did somewhat better with the S format [...] than the R format [...]; low-anxiety students did substantially better with the R format [...] than the S format [...]." (NEELY, u. a. 2002, 109) Zusammenfassend konnte nicht eindeutig nachgewiesen werden, dass die eine Form leichter, beziehungsweise bessere Ergebnisse lieferte, als die andere.

Eine weitere Studie beschäftigte sich mit der Frage "if there were differential rates of changing answers on test items according to the cognitive level of the items, and if the proportion of gains to losses across the items was different." (SMITH, u.a. 1979, 203) Items niedriger Ordnung wurden als Wissen und Verstehen, die höherer Ordnung als kognitive Prozesse wie Synthese, Analyse, Verwendung und Evaluation, definiert. Die Autoren gingen davon aus, dass die Probanden bei der ersten Art der Items entweder die Frage beantworten könnten oder nicht. Bei der zweiten Art würden die Befragten wohl überlegen, ob sie nicht einen voreiligen Fehler gemacht hätten. 157 Studenten hatten drei Multiple Choice Tests im Laufe eines Semesters bei denen insgesamt 56 Items niedriger Ordnung und 37 Items höherer Ordnung zur Analyse verwendet wurden. Die Studenten wurden dazu aufgefordert, falls sie Antworten während des Tests veränderten, dies so zu markieren, dass dem Prüfer ersichtlich wurde, welche die erste Antwort war. Zur Auswertung wurden folgende Kriterien entwickelt. Wurde eine falsche Antwort durchgestrichen und eine richtige stattdessen angekreuzt, zählt dies als Gewinn, umgekehrt als Verlust und von einer falschen zu einer falschen Antwort wurde als neutral eingeordnet. Beim letzten Test des Semesters wurden die Studenten zusätzlich befragt, inwiefern sie glaubten, eine Veränderung beim Ankreuzen der Antworten hätte Konsequenzen dahingehend, dass dies zu einer Punkteverminderung, Punkteerhöhung führen würde oder dies keinen Einfluss hätte. Es konnten bei dieser Studie Hinweise gefunden werden, dass mehr Antworten bei Items höherer Ordnung geändert wurden – die Anzahl dieser Items war niedriger – wobei sich jedoch insgesamt kein Unterschied in Gewinnen oder Verlusten bei den beiden Itemtypen zeigten (SMITH, u.a. 1979, 203ff). "Changing

answers on multiple-choice tests appears to be a rather common aspect of test taking.” (SMITH, u.a. 1979, 207) Obwohl sich zeigte, dass Änderungen förderlich waren, meinten die meisten Studenten, dies wirke sich negativ aus. Nur 7% waren der Ansicht, dass Änderungen die Wahrscheinlichkeit erhöhte, die richtige Antwort zu geben und 25% der Befragten gaben an, dass Änderungen keinen Unterschied machten (SMITH, u.a. 1979, 207).

Bei der Erforschung von drei verschiedenen Verfahren zur Distraktorenwahl konnte beobachtet werden, dass alle drei Methoden gleich valide und reliabel sind. Bei diesen handelte es sich um die “judgmental”, “frequency” und “discrimination” Methoden. Bei der Ersten nahm der Konstrukteur jene Distraktoren, die er für plausibel hielt, bei der Zweiten wurden von offenen Fragen die häufigsten falschen Antworten als Distraktoren herangezogen und bei der dritten Methode wurden die falschen Antworten bei offenen Fragen verwendet, die am besten zwischen “high and low score students” unterschieden (OWENS, u.a. 1970, 87f).

BUTLER und andere (2007, 273ff) untersuchten die Auswirkungen von Feedback auf das Lernen von Multiple Choice Tests. Dazu verwendeten sie zwei Arten der Rückmeldung, auf der einen Seite “standard versus answer-until-correct feedback” und auf der anderen Seite “immediate versus delayed feedback”. Beim “standard” Feedback wurde angezeigt, ob die gegebene Antwort richtig oder falsch war, beim “answer-until-correct” Feedback konnten die Studenten solange die Antworten wählen, bis sie die Richtige gefunden hatten, dabei mussten die Befragten überlegen, warum eine andere Antwort, als die von ihnen gewählte, richtig war. Beim “immediate feedback” bekamen die Studenten nach jeder Frage eine Rückmeldung, beim “delayed” eine verspätete. Das Ziel dieser Untersuchung war “to disentangle the effects of the type of feedback [...] and the timing of feedback [...] on long-term retention assessed on a delayed criterial test.” (BUTLER, u.a. 2007, 274) Dazu wurden zwei Experimente durchgeführt, bei denen alle Befragten “a series of brief prose passages” über pädagogische Themen und anschließend einen Multiple Choice Test bekamen. Zur Kontrolle wurden einige Fragen ohne Feedback gegeben, nur mit der Anweisung zu warten, bis die nächste Frage

erschien. Alle Studenten wurden am Ende des Multiple Choice Test mit anderen Distraktoraufgaben beschäftigt. Im ersten Experiment bekamen die Studenten einen Tag später die Fragen des Vortages und 8 zusätzliche Fragen in unterschiedlicher Reihenfolge vorgegeben und wurden aufgefordert mit einem Wort oder einem Satz zu antworten und gegebenenfalls zu raten. Die Autoren konnten einen Testeffekt, die Studenten schnitten am zweiten Tag besser ab, beobachten. Ebenso wurden Hinweise gefunden, dass das Feedback dafür unterstützend war. Das "answer-until-correct" Feedback stellte eine bessere Methode dar als das "standard" Feedback. Experiment zwei unterschied sich vom Ersten nur darin, dass der zweite Test eine Woche später stattfand. Auch in diesem konnten Testeffekte und die Hilfe vom Feedback für die Testleistung herausgefunden werden. Zusammenfassend konnten BUTLER und andere (2007, 279) aufzeigen, dass "delayed feedback produced better long term retention than immediate feedback" und dass es keinen Unterschied machte, welcher Typ von Feedback verwendet wurde. Feedback erhöhte die Testleistungen und "that students are motivated to fully process the feedback" (BUTLER, u. a. 2007, 280).

II.7 Schülerurteile

Subjektiv ist das, was von Personen erlebt wird, wobei es kein richtig oder falsch gibt. Der objektive Unterricht kann von verschiedenen Schülern sehr unterschiedlich wahrgenommen werden. „Gute Schüler werden den Unterrichtsstoff als leichter wahrnehmen als schlechte; Jungen erleben möglicherweise die Kameradschaft in der Klasse anders als die Mädchen.“ (DREESMANN 1982, 22)

In der Evaluation von Unterricht ist der Stellenwert von Schülerurteilen umstritten. Auf der einen Seite befürchten Praktiker, dass sich Schüler von persönlichen Vorlieben, Abneigungen, Erwartungshaltungen, Wunschvorstellungen, sowie Entschuldigungsstrategien für schlechte Leistungen in ihrem Urteil leiten lassen und dieses somit verzerrt wird. Auf der anderen Seite können Schülerurteile durchaus herangezogen werden, da Schüler – wie sonst niemand – mit den Lehrer-Verhaltensweisen konfrontiert sind und ihr Erfolg auch von der Wahrnehmung der Lehrerpersönlichkeit abhängt (STOLZ 1997, 124). Schüleraussagen spiegeln nicht nur persönliche Präferenzen und Vorurteile einzelner Schüler wieder, sondern auch Erwartungsmuster aufgrund schulischer Sozialisation (GERSTENMAIER 1975, 168f). „Die Bedeutung der Beliebtheit von Lehrern und ihrer psychologischen Determinanten für die Sozialisation von Schülern im Sinne der Verinnerlichung gesellschaftlicher Wertvorstellungen wie Leistungsbereitschaft und Verhaltenskonformität“ zeigt, „daß durch die Art der Schüler-Lehrer-Beziehung und die Lernziele selbst [...] ein bestimmter Druck zu Beliebtheitsurteilen besteht.“ (GERSTENMAIER 1975, 10) Beliebte Lehrer werden zu Modellpersonen und Urteile über diese zu Bewertungen der curricularen Lernziele (GERSTENMAIER 1975, 10).

Einige Autoren vertreten die Ansicht, dass „subjektive Wahrnehmungsdaten den standardisierten Beobachtungsdaten vorzuziehen sind, da sich solche subjektiven Perzeptionen meist auf einen längeren Erfahrungszeitraum beziehen.“ (CLAUSEN 2002, 43) Die Aussagen von Schülern stützen sich also auf Erfahrungen über einen längeren Zeitraum. Zwar werden diese von Noten und der Einstellung zur

Schule beeinflusst, aber laut DITTON (2002, 264/281) nicht auf Kosten ihrer Aussagekraft.

Schülerurteile haben den Nachteil, dass die Wahrnehmung eines Schülers von seinem Leistungsstand gelenkt wird (CLAUSEN 2002, 44). MOOS und BABAD (1979, 1996; zit. nach CLAUSEN 2002, 44) konnten beobachten, dass Schüler mit guten Noten ihre Umwelt, beziehungsweise den Unterricht, positiver beurteilen als schlechte Schüler.

Bei der Wahrnehmung werden Umweltinformationen selektiv aufgenommen und bei der Informationsverarbeitung subjektiv gewichtet, sowie emotional eingefärbt, wodurch das persönliche Erleben konstruiert wird (CLAUSEN 2002, 81). Das heißt, es kann nicht vermieden werden, dass Urteile von Stimmungen beeinträchtigt werden (KLINGLER 1982, 53). Die individuelle Wahrnehmung wird von Zielen, Fähigkeiten, Bedürfnissen sowie von Erfahrungen, Einstellungen und Motiven, die soziale Wahrnehmung dagegen auch noch durch die Einstellungen und Ziele des Gegenübers beeinflusst (CLAUSEN 2002, 81).

Bei einer Befragung von Schülern zu verschiedenen Unterrichtsmethoden bekam FICHTEN (1993, 98f) sehr widersprüchliche Ergebnisse und Antworten von den Jugendlichen, die auf deren Lernerfahrungen und spezifischen Bedürfnislagen zurück zu führen waren. Er beschäftigte sich daraufhin näher mit der lerngeschichtlichen Komponente und konnte am Beispiel Deutschunterricht folgendes feststellen: Die von ihm verwendete Methode der Gruppenarbeit wurde in den Jahren zuvor und in anderen Unterrichtsfächern kaum praktiziert. Diese ungeübte Methode hinterließ bei den Schülern teilweise frustrierende und ermüdende Eindrücke, wodurch es auch zu Misserfolgserlebnissen kam. Daher musste erst methodische Kompetenz aufgebaut werden, um positive Erfahrungen bei den Schülern zu erzeugen und somit deren Bewertungen langfristig und nachhaltig zu beeinflussen.

Wie sich in diesem Kapitel zeigte, findet man in der Literatur hauptsächlich Stellungnahmen von Schülern zum Unterricht, jedoch nicht, inwiefern Schüler Schularbeiten oder Tests beurteilen.

II.8 Zusammenfassung

Für den nachfolgenden empirischen Teil dieser Arbeit werde ich nochmals die wichtigsten Punkte zusammenfassen. Multiple Choice Tests werden eine hohe Ökonomie und Objektivität nachgesagt. Immer wieder stellt die Frage, welche kognitiven Leistungen mittels dieses Formats gemessen werden können, einen Streitpunkt in der Literatur dar. Offene Fragen werden ebenso kontrovers dargestellt, daher habe ich schon in der Einleitung vorweg genommen, dass es mir um keinen Vergleich in Richtung besser oder schlechter geht. Beide Verfahren haben ihre Vor- und Nachteile und je nach dem was der Lehrer prüfen will, entscheidet er sich für eines dieser Formate.

Bei all den Studien die ich zu diesem Thema gelesen habe, ging es kaum um die Einschätzung von Schülern oder Studenten. Es fehlte mir die direkte Sichtweise, Einschätzung beziehungsweise Beurteilung vom Schüler selbst. Dies stellte anschließend auch meine Hauptforschungsfrage dar: Wie bewerten Schüler Multiple Choice Tests?

Aufgrund der Literatur und des mir zur Verfügung gestellten Datensatz (siehe Kapitel III) ergaben sich folgende Fragestellungen:

Wie wird der vorgegebene Multiple Choice Test von den Schülern eingeschätzt?
Bei welchem Prüfungsformat geben die Schüler an mehr lernen zu müssen?
Welche Form des Prüfens schätzen die befragten Schüler als schwieriger ein?
Welches Prüfungsformat bevorzugen die Befragten Schüler?

Diese Fragen werden genauer im nächsten Kapitel betrachtet, wobei lediglich Tendenzen der Befragten Schüler dargestellt werden und keine Verallgemeinerungen getroffen werden.

III. Empirischer Teil

III.1 Methode

Der Datensatz dieser Diplomarbeit wurde im Rahmen einer Dissertation (FÜHRER 2007) im Sinne eines Prä- und Postdesigns erhoben.

III.1.1 Erhebungsinstrumente

Über den Zeitraum von drei Schuljahren, 2005-2008, wurden Jugendliche von der zweiten bis zur fünften Schulstufe einer Höheren Technischen Lehranstalt in diversen Gegenständen mittels Multiple Choice Tests getestet. Mit diesen Tests wurden gleichzeitig alle Schüler einer Klasse über das jeweilige Stoffgebiet geprüft und aus den erhobenen Leistungen konnten dann die Noten der Schüler ermittelt werden (FÜHRER 2007, 1).

Die Multiple Choice Tests wurden am Computer in der Schule durchgeführt und anschließend elektronisch ausgewertet. Die Fragen beinhalteten Themen aus den jeweiligen Gegenständen (AINF, INIT, BUN, PRE)², wurden per Zufallsgenerator gereiht und erschienen jeweils einzeln am Bildschirm. D.h. erst nach erfolgter Bearbeitung einer Frage erschien die nächste; eine erneute Bearbeitung einer bereits beantworteten Frage war nicht mehr möglich. Nur im Schuljahr 05/06 waren alle Fragen am Bildschirm sichtbar und konnten durch scrollen bearbeitet werden, was der Prüfer in den Schuljahren darauf änderte (FÜHRER 2007, 133ff).

² AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRE ... Project Engineering

Ein Test beinhaltete zwischen 10 und 20 Multiple Choice Fragen, wobei zwischen 4 und 8 Antwortmöglichkeiten vorgegeben wurden (FÜHRER 2007, 155).

Vor jedem Test erfolgte eine standardisierte Testinstruktion durch den Prüfer, die auf folgende Punkte hinwies:

- „Es können mehrere Antworten richtig sein. Der Fall, dass keine oder alle Antworten richtig sind, kann ausgeschlossen werden
- Wenn 2 Antworten richtig sind und nur 1 richtige davon angekreuzt wird, wird die Hälfte der Punkte der Frage gerechnet. D.h. diese Frage wird NICHT als falsch gerechnet!
- Wenn man sich nicht sicher ist, ob man noch eine 2. Antwort ankreuzen soll, dann lieber nicht wählen, da die falsche von der richtigen „abgezogen“ wird.
- Sich nicht auf die Ratewahrscheinlichkeit verlassen; diese ist sehr gering!
- Trotzdem ist es besser eine Antwort anzukreuzen als keine, da man bei keiner angekreuzten Antwort max. 0 Punkte erzielt“ (FÜHRER 2007, 156).

Nach dieser Instruktion wurde der Test gestartet.

Die Standardisierung der Punkte wurde durch eine Gewichtung in dem Sinne durchgeführt, sodass alle richtigen Antwortalternativen summiert 100 Punkte ergaben. Die Punkteanzahl pro Frage wurde berechnet, indem die anteiligen Punkte den Gesamtscore 100 ergaben.

Dies wird nun kurz anhand eines Beispiels erläutert: Bei einem Multiple Choice Test mit 10 Fragen, konnten pro Frage 10 Punkte erreicht werden (100/10). Hatte nun ein Schüler alle richtigen Antworten bei einer Frage angekreuzt, erhielt er für diese 10 Punkte. Kreuzte ein Schüler von zwei richtigen Antworten nur eine an, erhielt er 5 Punkte, also die Hälfte der möglichen Punkte.

Nun wird noch der dritte Punkt der Instruktionen anhand eines Beispiels dargestellt: Kreuzte ein Schüler zwei richtige und eine falsche Antwort an, erhielt er für die Frage 3,33 Punkte ($3,33 + 3,33 - 3,33 = 3,33$) von 10 möglichen Punkten (FÜHRER 2007, 157). Die Gesamtpunkteanzahl entschied am Ende eines Tests über die Note des Schülers, wobei 0-50 Punkte ein Nicht genügend, 51-62

ein Genügend, 63-75 ein Befriedigend, 76-89 ein Gut und 90-100 ein Sehr gut ergaben (FÜHRER 2007, 111). Dieser Notenschlüssel wurde vom Prüfer festgelegt.

Um diese Multiple Choice Tests verbessern zu können, wurde ein Feedbackbogen – „ein Fragebogen ... der vom Schüler nach dem Test „ausgefüllt“ wird“ – vom Prüfer selbst entwickelt (FÜHRER 2007, 161). „Feedback ist die einfachste und schnellste Form der Evaluation.“ (MITTELSTÄDT 2006, 112)

Den Feedbackbogen änderte FÜHRER (2007, 166) im Laufe seiner Vortestungen, um Verständnisprobleme und dergleichen zu beheben. Für diese Arbeit wurden die Feedbackversionen 4 und 5 verwendet, die sich durch formale, nicht jedoch durch inhaltliche Kriterien unterschieden. Der Feedbackbogen beinhaltete 23 Fragen, unter anderem über den Test, den Unterricht und das Stoffgebiet, und ist im Anhang dieser Arbeit dargestellt (Appendix A). Diese 23 Fragen wurden nach dem Multiple Choice Test den Schülern über den Computer vorgegeben, wobei diese mittels „Radio Buttons“ nur jeweils eine Antwortmöglichkeit ankreuzen konnten (FÜHRER 2007, 161). Vor dem Fragebogen bekamen die Schüler, wie bei den Tests, eine standardisierte Instruktion, die auf folgende Punkte hinwies:

- „nach Möglichkeit alle Fragen beantworten
- so aufrichtig wie möglich beantworten
- wenn man sich absolut nicht entscheiden kann, dann besser keine Antwort wählen
- möglichst zügig (1. Impuls ist meist der richtige)
- es gibt keine richtigen oder falschen Antworten
- bei diesen Feedbacks auch den Namen auswählen³, damit Auswertungen bezüglich Testzeit und Ergebnis erfolgen können
- es wird nicht ausgewertet WER WAS beantwortet
- Voraussetzung, dass Fragen *verständlich* und *klar* sind, ist natürlich, dass dieses Stoffgebiet gelernt wurde“ (FÜHRER 2007, 164).

³ Die mir zur Verfügung gestellten Daten waren anonymisiert.

III.1.2 Datenerhebung

Die Daten wurden wie bereits erwähnt über drei Schuljahre von 2005 bis 2008 an einer Höheren Technischen Lehranstalt erhoben. In den Schuljahren 05/06 und 06/07 nahmen jeweils zwei zweite, eine dritte, eine vierte und eine fünfte Klasse und im Schuljahr 07/08 zwei zweite, eine dritte, zwei vierte und eine fünfte Klasse an den Testungen und Feedbacks teil (siehe auch FÜHRER 2007, 166, 171, 181). Insgesamt wurden über die drei Jahre 49 Tests in den Fächern „Angewandte Informatik“, „Betriebssysteme und Netzwerke“, „Industrielle Elektronik in der Informationstechnik“ und „Project Engineering“ durchgeführt (FÜHRER 2007, 66). Je nach Klasse gab es zwischen 2 und 5 Tests sowie korrespondierenden Feedbacks pro Schuljahr. Jeder Test, inklusive dem nachgeschalteten Feedback, dauerte eine Schulstunde, also 50 Minuten.

III.2 Auswertung der Ergebnisse

In diesem Kapitel werden neben einer konkreten Stichprobenbeschreibung der Umgang mit dem anonymisierten Datensatz und die konkreten Ergebnisse, die diese Untersuchung lieferte, dargestellt.

III.2.1 Stichprobenbeschreibung

An der Testung mit nachgeschaltetem Feedback nahmen im Laufe der drei Jahre 276 Schüler und 4 Schülerinnen im Alter zwischen 15 und 21 Jahren teil. Die genaue Altersverteilung wird in Tabelle 1 dargestellt, wobei das Alter der Schüler mit dem Monat Juni des jeweiligen Schuljahres, beginnend bei Juni 2006, errechnet wurde. *N* stellt die Anzahl der Schüler pro Schuljahr dar. Da es sich hier um eine Längsschnittsstudie handelte, ist es von Bedeutung darzustellen, wie viele Schüler über welchen Zeitraum getestet wurden. 195 Schüler nahmen nur in einem Schuljahr teil, 61 an zwei und 24 an allen drei Schuljahren, wobei von den 280 Schülern insgesamt 19 eine Klasse wiederholten.

Schuljahr	<i>N</i>	Median	Minimum	Maximum	Range
2007/2008	134	17,71	15,00	21,42	6,42
2006/2007	130	17,58	15,83	21,58	5,75
2005/2006	125	17,42	15,67	21,50	5,83

Tabelle 1: Alter der Schüler

Für die nachfolgende Auswertung sind die Noten der Schüler bedeutsam, da sie die erbrachte Testleistung der Schüler widerspiegeln.

In Tabelle 2 sind die Noten aller Tests über die drei Schuljahre angeführt, wobei ersichtlich wird, dass im Schuljahr 06/07 die meisten Tests durchgeführt wurden.

Schuljahr	Summe der Noten	Fehlende	1er	2er	3er	4er	5er
2007/2008	391	0	23	90	108	79	91
		0,00%	5,88%	23,02%	27,62%	20,20%	23,27%
2006/2007	508	1	21	89	113	145	139
		0,20%	4,14%	17,55%	22,29%	28,60%	27,42%
2005/2006	343	0	26	54	73	82	108
		0,00%	7,58%	15,74%	21,28%	23,91%	31,49%

Tabelle 2: Darstellung der Notenverteilung über alle drei Schuljahre

Die Noten in Tabelle 2 spiegeln nicht die Gesamtjahresnoten wider, da sich diese neben den einzelnen Testergebnissen auch aus der Mitarbeit zusammensetzten.

III.2.2 Datenaufbereitung

Neben dem Alter und den Noten wurden die Schuljahre, das Fach und die dazugehörigen Klassen, die einzelnen Punkte pro Multiple Choice Frage, sowie die Gesamtpunkteanzahl jedes Schülers pro Test erhoben. Vom Feedbackbogen wurden die Angaben bezüglich acht Fragen, die für diese Arbeit relevant sind, für die Auswertung herangezogen; es handelte sich um die Fragen 2, 7, 12, 13, 16, 17, 18 und 19, die im Folgenden dargestellt werden.

„ 2) Die Fragen waren 'einfach'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

7) Das Stoffgebiet ist 'einfach'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

12) Der Lernaufwand bei 'Multiple Choice Tests' ist 'niedrig'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

13) Der Lernaufwand bei 'offenen Fragen' (z.B. Matheschularbeit; keine vorgefertigten Antworten) ist 'hoch'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

16) Der Schwierigkeitsgrad der Tests ist 'höher' bei

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

17) Der Lernaufwand ist 'höher' bei

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

18) Vergessenskurve: Wobei 'hält' das Wissen 'länger'?

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

19) Welche Tests sind Ihnen 'lieber'?

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)“

(FÜHRER 2007, 258f).

III.3 Darstellung der Ergebnisse

Die statistische Auswertung erfolgte mittels SPSS (Version 15.0) und dem Programm Comprehensive Meta-Analysis (CMA) Version 2⁴.

III.3.1 Korrelationen

Zur Messung des Zusammenhangs zwischen der Testleistung und der Beantwortung des dazugehörigen Feedbacks wurde ein Testscore und ein Feedbackscore errechnet. Der Testscore für jeden Schüler stellte die Summe der Gesamtpunkte pro Schuljahr, pro Klasse und pro Fach dar. Der Feedbackscore bildete die Summe für jede Feedbackfrage pro Schuljahr, pro Klasse und pro Fach. Alle Antworten, die ein Schüler zum Beispiel in der 2. Klasse im Schuljahr 2005/2006 im Fach „Angewandte Informatik“ auf Frage 2 gab, wurden summiert und bildeten somit diesen Score. Mittels der beiden genannten Scores wurden Korrelationen im SPSS gerechnet, die nun näher erläutert werden.

Da der Testscore metrisch, das Feedback bei Frage 2, 7, 12 und 13 rangskaliert und bei Frage 16, 17, 18 und 19 dichotom war, wurden Korrelationen nach Spearman und Pearson gerechnet. Obwohl es sich bei den ersten vier Fragen um rangskalierte Daten handelte, wurden hier sowohl Korrelationen nach Spearman, als auch nach Pearson gerechnet. Mittels dieser Korrelationen konnte jedes Schuljahr einzeln nach Fach und Feedbackfrage getrennt untersucht werden. Dies würde bedeuten, dass es viele einzelne Zusammenhangsmaße innerhalb kleiner Stichproben gibt, die jedoch keine Tendenzen über alle Schüler zeigen würden, sondern nur über die einzelnen Klassen, nach Fächern getrennt.

⁴ Für Interessierte ist das Programm CMA unter www.meta-analysis.com zu finden.

Da mich bei meiner Untersuchung der Zusammenhang zwischen dem Testscore und dem Feedbackscore unabhängig vom Fach und von der Klasse interessierte, musste ich auf ein anderes statistisches Programm zurückgreifen, mit dem es möglich war, diese Einzelinformationen zu einer Gesamtinformation zusammenzufassen. CMA ist ein statistisches Programm, welches in der Lage ist, von einander unabhängige statistische Maßzahlen in den einzelnen Stichproben für den Fall vergleichbarer Treatments zu kombinieren und Daten zusammenzufassen (Effektstärke), (BORENSTEIN, HEDGES, u.a., 2005). In meiner Untersuchung war der Feedbackbogen über alle Testungen immer der gleiche. Effektstärken „sind standardisierte, vergleichbare Schätzwerte für das Ausmaß, in dem ein experimenteller Effekt, eine Beziehung oder ein Phänomen in einer Population vorhanden ist.“ (PLATH 2006, 476)

In der vorliegenden Untersuchung wird das Zusammenhangsmaß r , das formal Pearsons Korrelationskoeffizienten entspricht, verwendet.

Um herauszufinden, ob sich Spearman- und Pearsonkorrelation voneinander unterscheiden, wurde mittels CMA untersucht, ob die beiden Gesamtkorrelationen (Spearman und Pearson) hinsichtlich des Gesamteffekts signifikant differieren, wenn nicht, könnten Gesamteffektstärken nach Pearson ermittelt werden.

Im ersten Schritt wurden Korrelationen zwischen den einzelnen Feedbackfragen und den dazugehörigen Testscores nach Schuljahren, Klassen und Fach getrennt, berechnet. Im zweiten Schritt wurde überprüft, ob sich die Korrelationen bei den Feedbackfragen 2, 7, 12 und 13 unterscheiden. Diese Prüfung auf Homogenität wird in den nachfolgenden Tabellen dargestellt.

Bei den Tabellen 3-6 werden unter n die Anzahl der miteinander zur Gesamtkorrelation kombinierten Korrelationen und unter r die Höhe der aus den Einzelkorrelationen resultierende Gesamteffektstärke verstanden.

Fixed effect analysis	Anzahl der Stichproben	Effektstärke und 95% Intervall			Hypothesentest (2-seitig)		Heterogenität		
		r	KI_u^5	KI_o^6	z	p	Q	df	p
bivariat	17	,302	,209	,391	6,094	< ,001	19,497	16	,244
rang	17	,294	,199	,382	5,905	< ,001	23,979	16	,090
zwischen							0,018	1	,894

Tabelle 3: Unterschiede in den Gesamteffektstärken von r_{xy} und r' bei Feedbackfrage 2

Zeigte sich bei den Gesamteffekten der Korrelationen keine Heterogenität der Einzelkorrelationen, wurde die Gesamteffektstärke mittels Fixed Effects, die eine zugrunde liegende Homogenität der Einzelkorrelationen voraussetzt, berechnet. Im Falle einer zugrunde liegenden Heterogenität wäre eine Random Effects Berechnung zur Anwendung gekommen, die beobachtete, durch Moderatorvariable nicht erklärbare Mischverteilungen durch einen Korrekturfaktor berücksichtigt.

Wie sich in Tabelle 3 zeigt, war der Heterogenitätstest in Bezug auf Feedbackfrage 2 nicht signifikant ($Q = 0,018$; $df = 1$; $p = ,894$). Die Korrelationen r_{xy} und r' unterschieden sich nicht signifikant voneinander, daher konnten die weiteren Ergebnisse mit der Korrelation von Pearson r_{xy} berechnet werden.

⁵ KI_u ... Untere Grenze des Konfidenzintervalls

⁶ KI_o ... Obere Grenze des Konfidenzintervalls

Fixed effect analysis	Anzahl der Stichproben	Effektstärke und 95% Intervall			Hypothesentest (2-seitig)		Heterogenität		
		r	Kl_u	Kl_o	z	p	Q	df	p
bivariat	17	,309	,215	,396	6,226	< ,001	21,588	16	,157
rang	17	,274	,179	,364	5,484	< ,001	24,938	16	,071
zwischen							0,275	1	,600

Tabelle 4: Unterschiede in den Gesamteffektstärken von r_{xy} und r' bei Feedbackfrage 7

Wie sich in Tabelle 4 zeigt, war der Heterogenitätstest in Bezug auf Feedbackfrage 7 nicht signifikant ($Q = 0,275$; $df = 1$; $p = ,600$). Die Korrelationen r_{xy} und r' unterschieden sich nicht signifikant voneinander, daher konnten die weiteren Ergebnisse mit der Korrelation von Pearson r_{xy} berechnet werden.

Fixed effect analysis	Anzahl der Stichproben	Effektstärke und 95% Intervall			Hypothesentest (2-seitig)		Heterogenität		
		r	Kl_u	Kl_o	z	p	Q	df	p
bivariat	17	,139	,040	,236	2,736	,006	19,604	16	,239
rang	17	,135	,035	,232	2,651	,008	21,731	16	,152
zwischen							0,004	1	,952

Tabelle 5: Unterschiede in den Gesamteffektstärken von r_{xy} und r' bei Feedbackfrage 12

Wie sich in Tabelle 5 zeigt, war der Heterogenitätstest in Bezug auf Feedbackfrage 12 nicht signifikant ($Q = 0,004$; $df = 1$; $p = ,952$). Die Korrelationen r_{xy} und r' unterschieden sich nicht signifikant voneinander, daher konnten die weiteren Ergebnisse mit der Korrelation von Pearson r_{xy} berechnet werden.

Fixed effect analysis	Anzahl der Stichproben	Effektstärke und 95% Intervall			Hypothesentest (2-seitig)		Heterogenität		
		r	Kl_u	Kl_o	z	p	Q	df	p
bivariat	17	,149	,050	,245	2,929	,003	17,032	16	,384
rang	17	,106	,006	,204	2,082	,037	12,329	16	,721
zwischen							0,358	1	,550

Tabelle 6: Unterschiede in den Gesamteffektstärken von r_{xy} und r' bei Feedbackfrage 13

Wie sich in Tabelle 6 zeigt, war der Heterogenitätstest in Bezug auf Feedbackfrage 13 nicht signifikant ($Q = 0,358$; $df = 1$; $p = ,550$). Die Korrelationen r_{xy} und r' unterschieden sich nicht signifikant voneinander, daher konnten die weiteren Ergebnisse mit der Korrelation von Pearson r_{xy} berechnet werden.

Da sich keine Unterschiede zwischen den Korrelationen ergaben, konnte nun mittels CMA aus den einzelnen Korrelationen eine Gesamtkorrelation der einzelnen Fragen, über alle Schuljahre und Klassen errechnet werden. Insgesamt wurde aus 17 Korrelationen pro Feedbackfrage mit korrespondierendem Testscore eine Gesamtkorrelation r_G , beziehungsweise eine Gesamteffektstärke pro Feedbackfrage mit korrespondierendem Testscore errechnet.

Der Zusammenhang zwischen den Feedbackfragen 2 ($r_G = ,302$; $p < ,001$), 7 ($r_G = ,309$; $p < ,001$), 12 ($r_G = ,139$; $p = ,006$), 13 ($r_G = ,149$; $p = ,003$), 16 ($r_G = ,144$; $p = ,005$), 17 ($r_G = ,196$; $p < ,001$) und 18 ($r_G = ,131$; $p = ,010$) und den dazugehörigen Testleistungen war signifikant. Nur bei Frage 19 ($r_G = ,066$; $p = ,195$) konnte kein signifikanter Zusammenhang festgestellt werden. Im Folgenden sind die Berechnungen der Gesamteffektstärken tabellarisch dargestellt.

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,388**	,113	,607	,007
2007/2008	3A/INIT	26	,276	-,125	,599	,174
2007/2008	4A/PRE	19	,239	-,241	,625	,330
2007/2008	4AF/BUN	18	,418	-,061	,740	,085
2007/2008	5A/INIT	21	-,056	-,476	,385	,812
2006/2007	2A/AINF	25	-,151	-,515	,260	,475
2006/2007	2B/AINF	32	,519**	,208	,735	,002
2006/2007	3A/INIT	24	,402	-,002	,693	,051
2006/2007	3A/PRE	25	,102	-,305	,478	,631
2006/2007	4AF/BUN	22	,382	-,047	,692	,079
2006/2007	5A/INIT	24	,293	-,125	,623	,167
2005/2006	2A/AINF	26	,575**	,241	,787	,002
2005/2006	2B/AINF	28	,470*	,118	,717	,011
2005/2006	3A/INIT	24	,059	-,353	,452	,787
2005/2006	4AF/BUN	19	,248	-,232	,631	,311
2005/2006	5A/INIT	26	,496**	,134	,741	,009
2005/2006	5A/PRE	26	,031	-,361	,413	,882
Gesamteffektstärke		<i>N</i> = 432	,302***	,209	,391	< ,001

Tabelle 7: Zusammenhang zwischen der Testleistung und der Feedbackfrage 2⁷

⁷ * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRES ... Project Engineering

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,328*	,045	,562	,024
2007/2008	3A/INIT	26	,360	-,032	,656	,071
2007/2008	4A/PRE	19	,325	-,152	,679	,177
2007/2008	4AF/BUN	18	,325	-,167	,688	,192
2007/2008	5A/INIT	21	,225	-,229	,599	,331
2006/2007	2A/AINF	25	-,091	-,469	,315	,669
2006/2007	2B/AINF	32	,532***	,225	,743	,001
2006/2007	3A/INIT	24	,181	-,240	,545	,402
2006/2007	3A/PRE	25	,317	-,089	,633	,124
2006/2007	4AF/BUN	22	,457*	,044	,737	,031
2006/2007	5A/INIT	24	,089	-,326	,475	,683
2005/2006	2A/AINF	26	,557**	,216	,777	,003
2005/2006	2B/AINF	28	,583***	,268	,785	,001
2005/2006	3A/INIT	24	,231	-,190	,580	,281
2005/2006	4AF/BUN	19	-,204	-,602	,276	,408
2005/2006	5A/INIT	26	,508**	,150	,748	,007
2005/2006	5A/PRE	26	,007	-,381	,393	,973
Gesamteffektstärke		<i>N</i> = 432	,309***	,215	,396	< ,001

Tabelle 8: Zusammenhang zwischen der Testleistung und der Feedbackfrage 7⁸

⁸ * * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRE ... Project Engineering

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,123	-,170	,396	,412
2007/2008	3A/INIT	26	,187	-,216	,536	,364
2007/2008	4A/PRE	19	,359	-,114	,699	,133
2007/2008	4AF/BUN	18	,143	-,347	,572	,577
2007/2008	5A/INIT	21	,190	-,263	,575	,414
2006/2007	2A/AINF	25	-,419*	-,698	-,029	,036
2006/2007	2B/AINF	32	,420*	,084	,670	,016
2006/2007	3A/INIT	24	,199	-,222	,558	,355
2006/2007	3A/PRE	25	-,164	-,525	,247	,438
2006/2007	4AF/BUN	22	,084	-,350	,488	,714
2006/2007	5A/INIT	24	,092	-,323	,478	,672
2005/2006	2A/AINF	26	,489**	,125	,737	,010
2005/2006	2B/AINF	28	,201	-,186	,534	,308
2005/2006	3A/INIT	24	,205	-,216	,562	,341
2005/2006	4AF/BUN	19	-,096	-,527	,375	,700
2005/2006	5A/INIT	26	,258	-,144	,587	,206
2005/2006	5A/PRE	26	-,070	-,445	,326	,737
Gesamteffektstärke		<i>N</i> = 432	,139**	,040	,236	,006

Tabelle 9: Zusammenhang zwischen der Testleistung und der Feedbackfrage 12⁹

⁹ * * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRE ... Project Engineering

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,235	-,056	,489	,112
2007/2008	3A/INIT	26	,307	-,091	,621	,128
2007/2008	4A/PRE	19	-,097	-,528	,374	,697
2007/2008	4AF/BUN	18	,328	-,164	,689	,187
2007/2008	5A/INIT	21	-,033	-,458	,404	,889
2006/2007	2A/AINF	25	,219	-,193	,565	,296
2006/2007	2B/AINF	32	,425*	,090	,674	,015
2006/2007	3A/INIT	24	,096	-,320	,481	,659
2006/2007	3A/PRE	25	,023	-,376	,414	,914
2006/2007	4AF/BUN	22	,092	-,343	,494	,688
2006/2007	5A/INIT	24	-,133	-,509	,286	,540
2005/2006	2A/AINF	26	,497**	,136	,742	,009
2005/2006	2B/AINF	28	,306	-,076	,609	,114
2005/2006	3A/INIT	24	-,013	-,414	,392	,952
2005/2006	4AF/BUN	19	-,248	-,631	,232	,311
2005/2006	5A/INIT	26	,136	-,265	,497	,512
2005/2006	5A/PRE	26	-,105	-,473	,294	,613
Gesamteffektstärke		<i>N</i> = 432	,149**	,050	,245	,003

Tabelle 10: Zusammenhang zwischen der Testleistung und der Feedbackfrage 13¹⁰

¹⁰ * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRE ... Project Engineering

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,071	-,221	,351	,637
2007/2008	3A/INIT	26	,307	-,091	,621	,128
2007/2008	4A/PRE	19	,359	-,114	,699	,133
2007/2008	4AF/BUN	18	,408	-,073	,735	,093
2007/2008	5A/INIT	21	,139	-,311	,538	,553
2006/2007	2A/AINF	25	,080	-,325	,461	,707
2006/2007	2B/AINF	32	,313	-,040	,597	,081
2006/2007	3A/INIT	24	,130	-,289	,507	,549
2006/2007	3A/PRE	25	-,170	-,530	,241	,421
2006/2007	4AF/BUN	22	,401	-,025	,704	,064
2006/2007	5A/INIT	24	-,054	-,448	,357	,804
2005/2006	2A/AINF	26	,390*	,003	,675	,048
2005/2006	2B/AINF	28	,220	-,167	,548	,263
2005/2006	3A/INIT	24	,121	-,297	,500	,577
2005/2006	4AF/BUN	19	-,017	-,468	,441	,946
2005/2006	5A/INIT	26	,116	-,284	,482	,576
2005/2006	5A/PRE	26	-,308	-,621	,090	,127
Gesamteffektstärke		<i>N</i> = 432	,144**	,044	,241	,005

Tabelle 11: Zusammenhang zwischen der Testleistung und der Feedbackfrage 16¹¹

¹¹ * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRE ... Project Engineering

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,254	-,036	,504	,085
2007/2008	3A/INIT	26	,429*	,050	,700	,028
2007/2008	4A/PRE	19	,227	-,253	,618	,355
2007/2008	4AF/BUN	18	,522*	,073	,795	,025
2007/2008	5A/INIT	21	,161	-,291	,554	,491
2006/2007	2A/AINF	25	,094	-,313	,472	,658
2006/2007	2B/AINF	32	,364*	,018	,632	,040
2006/2007	3A/INIT	24	,020	-,387	,420	,927
2006/2007	3A/PRE	25	,037	-,363	,426	,862
2006/2007	4AF/BUN	22	,371	-,060	,685	,089
2006/2007	5A/INIT	24	-,021	-,421	,386	,923
2005/2006	2A/AINF	26	,317	-,080	,627	,115
2005/2006	2B/AINF	28	,209	-,178	,540	,289
2005/2006	3A/INIT	24	,118	-,300	,498	,587
2005/2006	4AF/BUN	19	,189	-,290	,592	,444
2005/2006	5A/INIT	26	,048	-,346	,427	,818
2005/2006	5A/PRE	26	-,107	-,475	,292	,606
Gesamteffektstärke		<i>N</i> = 432	,196***	,098	,290	< ,001

Tabelle 12: Zusammenhang zwischen der Testleistung und der Feedbackfrage 17¹²

¹² * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

PRE ... Project Engineering

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,071	-,221	,351	,637
2007/2008	3A/INIT	26	,482*	,116	,733	,012
2007/2008	4A/PRE	19	,254	-,226	,635	,299
2007/2008	4AF/BUN	18	,226	-,269	,627	,373
2007/2008	5A/INIT	21	,157	-,295	,551	,502
2006/2007	2A/AINF	25	,033	-,367	,423	,877
2006/2007	2B/AINF	32	,349*	,000	,622	,050
2006/2007	3A/INIT	24	-,198	-,557	,223	,358
2006/2007	3A/PRE	25	,361	-,040	,662	,076
2006/2007	4AF/BUN	22	-,102	-,502	,334	,655
2006/2007	5A/INIT	24	-,367	-,671	,043	,078
2005/2006	2A/AINF	26	,370	-,020	,662	,062
2005/2006	2B/AINF	28	,212	-,175	,542	,282
2005/2006	3A/INIT	24	,260	-,160	,600	,223
2005/2006	4AF/BUN	19	-,205	-,603	,275	,406
2005/2006	5A/INIT	26	,241	-,161	,575	,238
2005/2006	5A/PRE	26	-,134	-,496	,267	,518
Gesamteffektstärke		<i>N</i> = 432	,131**	,031	,228	,010

Tabelle 13: Zusammenhang zwischen der Testleistung und der Feedbackfrage 18¹³

¹³ * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

Schuljahr	Klasse/Fach	<i>n</i>	<i>r</i>	Untere Grenze <i>KI</i>	Obere Grenze <i>KI</i>	<i>p</i>
2007/2008	2A_2B/AINF	47	,027	-,262	,312	,858
2007/2008	3A/INIT	26	,431*	,052	,701	,027
2007/2008	4A/PRE	19	,118	-,355	,543	,635
2007/2008	4AF/BUN	18	,308	-,186	,677	,218
2007/2008	5A/INIT	21	,120	-,329	,525	,609
2006/2007	2A/AINF	25	,064	-,340	,448	,764
2006/2007	2B/AINF	32	,160	-,200	,482	,385
2006/2007	3A/INIT	24	-,276	-,611	,143	,194
2006/2007	3A/PRE	25	,267	-,143	,599	,199
2006/2007	4AF/BUN	22	-,129	-,522	,309	,572
2006/2007	5A/INIT	24	-,288	-,619	,131	,174
2005/2006	2A/AINF	26	,180	-,223	,530	,383
2005/2006	2B/AINF	28	,004	-,370	,377	,984
2005/2006	3A/INIT	24	-,096	-,418	,320	,659
2005/2006	4AF/BUN	19	-,141	-,559	,335	,570
2005/2006	5A/INIT	26	,182	-,221	,532	,377
2005/2006	5A/PRE	26	,127	-,274	,490	,540
Gesamteffektstärke		<i>N</i> = 432	,066	-,034	,165	,195

Tabelle 14: Zusammenhang zwischen der Testleistung und der Feedbackfrage 19¹⁴

PRE ... Project Engineering

¹⁴ * = $p \leq ,050$; ** = $p \leq ,010$; *** = $p \leq ,001$

n ... Anzahl der Schüler pro Korrelation

N ... Summe der Schüler bei allen Korrelationen

KI ... Konfidenzintervall, $\alpha = ,050$

AINF ... Angewandte Informatik

INIT ... Industrielle Elektronik in der Informationstechnik

BUN ... Betriebssysteme und Netzwerke

III.3.2 Allgemeines lineares Modell für Messwiederholungen

Mit dem allgemeinen linearen Modell (ALM) wurden die Veränderungen der Rückmeldungen beim Feedback über die Zeit im ersten absolvierten Schuljahr der Schüler berechnet.

Um eine Vergleichbarkeit der Daten zu gewährleisten wurde wie folgt vorgegangen: Es wurde jeweils das erste Schuljahr, in dem ein Schüler an der Testung teilnahm, zur Auswertung herangezogen, um sicherzustellen, dass keine Testerfahrungseffekte auftraten. Die Berechnung des ALMs erfolgte klassenweise, da es pro Klasse unterschiedlich viele Testzeitpunkte gab. Der Testplan über alle drei Schuljahre ist im Anhang dieser Arbeit zu finden (Appendix B).

Die Tabelle 15 zeigt, welche Klassen zur Auswertung des ALMs herangezogen werden konnten.

Schuljahr	Klasse	Testzeitpunkte
2007/2008	4AF	3
2006/2007	2A	3
2006/2007	2B	4
2006/2007	4AF	4
2005/2006	2B	3
2005/2006	3A	3
2005/2006	5A	4 ¹⁵

Tabelle 15: Klassen, die für Messwiederholung in Frage kommen

PRE ... Project Engineering

¹⁵ Bei der 5A wurden die Schüler in zwei Fächern zu jeweils 2 Zeitpunkten mittels des Feedbacks befragt. Da es sich jedoch um ein und dieselben Schüler handelte, wurden die Testzeitpunkte nach dem Datum der Testung zusammengefasst.

Zur Berechnung wurden nur die Klassen verwendet, die an mehr als zwei Testungen teilnahmen, sowie die Feedbackfragen 2, 7, 12 und 13, da es sich bei den restlichen Fragen um dichotome Items handelte, deren Auswertung mittels ALM nicht sinnvoll war.

Aufgrund der Tatsache, dass auf der einen Seite nicht alle Schüler an jeder Testung in einem Schuljahr teilnahmen und auf der anderen Seite nicht alle Feedbackbogen ganz, beziehungsweise nur teilweise ausgefüllt wurden, variierte die Anzahl der Personen N bei den unten angeführten Ergebnissen.

Als Faktor wurde die Anzahl der Testzeitpunkte gewählt und als Innersubjektvariablen die einzelnen Feedbackfragen. Bei den folgenden Ergebnissen wurde ein Signifikanzniveau von $\alpha = ,050$ angenommen.

Schuljahr 2007/2008

- **4AF:** In dieser Klasse konnten bei den Fragen 2 ($N = 12$; $F = 1,505$; $df = 2$; $p = ,244$), 7 ($N = 12$; $F = 1,000$; $df = 2$; $p = ,384$) und 13 ($N = 11$; $F = 1,579$; $df = 2$; $p = ,231$) keine signifikanten Unterschiede bezüglich der Beantwortung über die drei Zeitpunkte festgestellt werden. Bei Frage 12 ($N = 11$; $F = 5,110$; $df = 2$; $p = ,016$) unterschieden sich die Antworten signifikant über die Zeit. Bei Frage 12 wurden $R = 33,8\%$ der Varianz durch die Testzeitpunkte aufgeklärt.

Schuljahr 2006/2007

- **2A:** Bei der Beantwortung der Feedbackfragen 2 ($N = 19$; $F = 6,429$; $df = 2$; $p = ,004$), 7 ($N = 17$; $F = 5,725$; $df = 2$; $p = ,007$) und 12 ($N = 18$; $F = 3,702$; $df = 2$; $p = ,035$) ergab sich ein signifikanter Unterschied zwischen den drei Zeitpunkten. Bei der Frage 13 ($N = 17$; $F = ,449$; $df = 2$; $p = ,642$) ergab sich kein signifikanter Unterschied bezüglich der Beantwortung der drei Feedbackbögen. Bei Frage 2 wurden $R = 26,3\%$, bei Frage 7 $R = 26,4\%$ und bei Frage 12 $R = 17,9\%$ der Varianz durch die Testzeitpunkte aufgeklärt.

- **2B:** Bei den Fragen 2 ($N = 22$; $F = 2,968$; $df = 3$; $p = ,039$) und 7 ($N = 22$; $F = 3,449$; $df = 3$; $p = ,022$) konnte eine signifikante Veränderung im Antwortverhalten über die vier Testzeitpunkte festgestellt werden, im Gegensatz zu den Fragen 12 ($N = 22$; $F = 1,990$; $df = 3$; $p = ,188$) und 13 ($N = 22$; $F = ,676$; $df = 3$; $p = ,570$), wo keine signifikante Veränderung sichtbar wurde. Bei Frage 2 wurden $R = 12,4\%$ und bei Frage 7 $R = 14,1\%$ der Varianz durch die Testzeitpunkte aufgeklärt.
- **4AF:** Bei der Beantwortung der Feedbackfragen 2 ($N = 14$; $F = ,703$; $df = 3$; $p = ,556$), 7 ($N = 13$; $F = ,188$; $df = 3$; $p = ,904$), 12 ($N = 12$; $F = ,947$; $df = 3$; $p = ,429$) und 13 ($N = 11$; $F = ,379$; $df = 3$; $p = ,769$) ergab sich kein signifikanter Unterschied zwischen den vier Zeitpunkten.

Schuljahr 2005/2006

- **2B:** In dieser Klasse ergab sich bezüglich der Beantwortung aller 4 Fragen ein signifikanter Unterschied zwischen den drei Zeitpunkten die nun dargestellt werden: Frage 2 ($N = 22$; $F = 6,720$; $df = 2$; $p = ,003$), Frage 7 ($N = 21$; $F = 5,972$; $df = 2$; $p = ,005$), Frage 12 ($N = 21$; $F = 16,831$; $df = 2$; $p < ,001$) und Frage 13 ($N = 21$; $F = 4,654$; $df = 2$; $p = ,015$). Bei Frage 2 wurden $R = 24,2\%$, bei Frage 7 $R = 23,0\%$, bei Frage 12 $R = 45,7\%$ und bei Frage 13 $R = 18,9\%$ der Varianz durch die Testzeitpunkte aufgeklärt.
- **3A:** Bei den Fragen 2 ($N = 17$; $F = 2,545$; $df = 2$; $p = ,094$) und 13 ($N = 16$; $F = ,789$; $df = 2$; $p = ,463$) konnte keine signifikante Veränderung im Antwortverhalten über die drei Testzeitpunkte festgestellt werden, im Gegensatz zu den Fragen 7 ($N = 18$; $F = 13,890$; $df = 2$; $p < ,001$) und 12 ($N = 16$; $F = 20,065$; $df = 2$; $p < ,001$), wo eine signifikante Veränderung zu beobachten war. Bei Frage 7 wurden $R = 45,0\%$ und bei Frage 12 $R = 57,2\%$ der Varianz durch die Testzeitpunkte aufgeklärt.

- **5A:** Bei der Beantwortung der Feedbackfragen 2 ($N = 8$; $F = 5,327$; $df = 3$; $p = ,007$), 7 ($N = 9$; $F = 8,038$; $df = 3$; $p = ,001$) und 12 ($N = 7$; $F = 9,458$; $df = 3$; $p = ,001$) ergab sich ein signifikanter Unterschied zwischen den vier Zeitpunkten. Bei der Frage 13 ($N = 7$; $F = 2,867$; $df = 3$; $p = ,065$) ergab sich kein signifikanter Unterschied bezüglich der Beantwortung über die Zeit. Bei Frage 2 wurden $R = 43,2\%$, bei Frage 7 $R = 50,1\%$ und bei Frage 12 $R = 61,2\%$ der Varianz durch die Testzeitpunkte aufgeklärt.

III.3.3 t-Test bei abhängigen Stichproben

In Tabelle 16 werden die Klassen dargestellt, die im ersten Schuljahr zwei Multiple Choice Tests mit nachgeschaltetem Feedback, jeweils eine Testung pro Semester, hatten. Beim vorliegenden T-Test für abhängige Stichproben werden, wie beim ALM, die Veränderungen der Antworten pro Feedbackfrage für jedes Schuljahr und jede Klasse getrennt berechnet.

Ein hoher Mittelwert bei den Feedbackfragen 2, 7, 12, 13 entspricht der Antwortkategorie „trifft nicht zu“ beziehungsweise „trifft eher nicht zu“ und ein niedriger Mittelwert der Kategorie „trifft zu“ beziehungsweise „trifft eher zu“. Bei den folgenden Ergebnissen wurde ein Signifikanzniveau von $\alpha = ,050$ angenommen.

Schuljahr	Klasse	Testzeitpunkte
2007/2008	2A	2
2007/2008	2B	2
2006/2007	5A	2
2005/2006	2A	2
2005/2006	4AF	2

Tabelle 16: mittels T-Test ausgewertete Daten

Schuljahr 2007/2008

- **2A:** In dieser Klasse konnten keine signifikanten Veränderungen im Antwortverhalten in Bezug auf Frage 2 ($N = 20$; $MW_1 = 2,350$; $MW_2 = 2,650$; $T = -1,371$; $df = 19$; $p = ,186$), 7 ($N = 20$; $MW_1 = 2,900$; $MW_2 = 2,800$; $T = ,698$; $df = 19$; $p = ,494$) und 12 ($N = 20$; $MW_1 = 2,600$; $MW_2 = 2,650$; $T = -203$; $df = 19$; $p = ,841$) beobachtet werden. Bei Frage 13 ($N = 20$; $MW_1 = 3,300$; $MW_2 = 2,850$; $T = 2,269$; $df = 19$; $p = ,035$) konnte jedoch eine signifikante Veränderung gefunden werden.
- **2B:** Hier konnte bei Frage 2 ($N = 23$; $MW_1 = 1,957$; $MW_2 = 2,348$; $T = -2,105$; $df = 22$; $p = ,047$) ein signifikanter Unterschied zwischen den Antworten gefunden werden. Bei den Fragen 7 ($N = 23$; $MW_1 = 2,174$; $MW_2 = 2,000$; $T = 1,073$; $df = 22$; $p = ,295$), 12 ($N = 23$; $MW_1 = 2,044$; $MW_2 = 2,044$; $T = ,000$; $df = 22$; $p = 1,000$) und 13 ($N = 23$; $MW_1 = 3,087$; $MW_2 = 3,087$; $T = ,000$; $df = 22$; $p = 1,000$) wurde kein signifikanter Unterschied gefunden.

Schuljahr 2006/2007

- **5A:** Bei den folgenden drei Fragen konnte kein signifikanter Unterschied zwischen der Beantwortung der Fragen zum ersten und zum zweiten Zeitpunkt gefunden werden: 2 ($N = 20$; $MW_1 = 2,100$; $MW_2 = 1,750$; $T = 1,789$; $df = 19$; $p = ,090$), 7 ($N = 19$; $MW_1 = 2,737$; $MW_2 = 2,684$; $T = ,294$; $df = 18$; $p = ,772$) und 13 ($N = 20$; $MW_1 = 3,050$; $MW_2 = 2,850$; $T = ,890$; $df = 19$; $p = ,385$). Bei der Frage 12 ($N = 20$; $MW_1 = 1,950$; $MW_2 = 2,750$; $T = -3,107$; $df = 19$; $p = ,006$) unterschieden sich die Antworten signifikant.

Schuljahr 2005/2006

- **2A:** Bei den Fragen 2 ($N = 22$; $MW_1 = 2,046$; $MW_2 = 2,727$; $T = -3,382$; $df = 21$; $p = ,003$) und 12 ($N = 23$; $MW_1 = 2,087$; $MW_2 = 3,261$; $T = -4,013$; $df = 22$; $p = ,001$) konnte eine signifikante Veränderung im Antwortverhalten zwischen den zwei Zeitpunkten festgestellt werden. Bei den anderen beiden, 7 ($N = 23$; $MW_1 = 2,652$; $MW_2 = 3,087$; $T = -1,931$; $df = 22$; $p = ,066$) und 13 ($N = 22$; $MW_1 = 3,546$; $MW_2 = 3,773$; $T = -1,226$; $df = 21$; $p = ,234$), konnte keine signifikante Veränderung beobachtet werden.
- **4AF:** In dieser Klasse konnte für die Feedbackfragen 2 ($N = 17$; $MW_1 = 1,765$; $MW_2 = 3,294$; $T = -6,685$; $df = 16$; $p < ,001$) und 12 ($N = 18$; $MW_1 = 2,222$; $MW_2 = 3,000$; $T = -2,439$; $df = 17$; $p = ,026$) ein signifikanter Unterschied zwischen dem Antwortverhalten zwischen den zwei Zeitpunkten festgestellt werden. Bei Frage 7 ($N = 16$; $MW_1 = 2,250$; $MW_2 = 2,813$; $T = -1,649$; $df = 15$; $p = ,120$) und 13 ($N = 17$; $MW_1 = 3,588$; $MW_2 = 3,706$; $T = -,566$; $df = 16$; $p = ,579$) zeigte sich kein signifikanter Unterschied.

III.4 Beschreibung der Ergebnisse

In diesem Teil der Arbeit werden nun die Ergebnisse näher beschrieben und erläutert.

III.4.1 Zusammenhang zwischen der Testleistung und den Feedbackfragen

Korrelationen können „als eine Beschreibung einer allgemeinen Tendenz“ angesehen werden (SCANNELL, TRACY 1977, 189). Das heißt, dass die geschilderten Ergebnisse Tendenzen darstellen, zu denen die Befragten Schüler neigen.

Bei der Untersuchung, ob es einen Zusammenhang zwischen der Testleistung und den einzelnen Feedbackfragen gibt, wurden folgende Ergebnisse gefunden: Bei der Prüfung, ob ein Zusammenhang zwischen der Wahrnehmung der Frage als einfach und dem erzielten Testscore besteht, konnte ein hoch signifikanter Zusammenhang ($r = ,302$; $p < ,001$) zwischen hoher Testleistung und Wahrnehmung der Fragen als einfach beobachtet werden.

Bei Feedbackfrage 7 konnte ein hoch signifikanter Zusammenhang ($r = ,309$; $p < ,001$) zwischen hoher Testleistung und Wahrnehmung des Stoffgebietes als einfach festgestellt werden.

Die Frage nach dem Zusammenhang zwischen der Testleistung und dem wahrgenommenen Lernaufwand ergab ein niedriges signifikantes Ergebnis. Je höher also die Testleistung ist, desto niedriger wird der Lernaufwand für Multiple Choice Tests eingeschätzt ($r = ,139$; $p = ,006$).

Weiters wurde zum Vorschein gebracht, dass je höher die Testleistung, desto höher wird der Lernaufwand für offene Fragen eingeschätzt ($r = ,149$; $p = ,003$).

Bei der Feedbackfrage 16 wurde ein positiver Zusammenhang ($r = ,144$; $p = ,005$) zwischen der erzielten Testleistung und den wahrgenommenen

Schwierigkeitsgrad gefunden; Schüler die beim Test besser abschneiden, also eine höhere Punkteanzahl erzielen, nehmen offene Fragen als schwieriger war.

Schüler die im Multiple Choice Test besser abschneiden, schätzen den Lernaufwand bei offenen Fragen höher ein, und dass das Wissen bei Tests mit offenen Fragen länger hält ($r = ,196$; $p < ,001$ / $r = ,131$; $p = ,010$).

Kein Zusammenhang konnte zwischen der Präferenz für eines der beiden Formate und dem Testergebnis festgestellt werden ($r = ,066$; $p = ,193$), es konnte also nicht gezeigt werden, ob Schüler die gut im Test abschneiden, eines der beiden Formate bevorzugen.

III.4.2 Veränderungen über die Zeit, bei mehr als zwei Testzeitpunkten

Im Schuljahr 2007/2008 wurden in der 4AF die Fragen des Tests sowie das Stoffgebiet über die Zeit als eher einfach eingeschätzt. In diesem Schuljahr wurde die Frage, ob der Lernaufwand bei offenen Fragen hoch ist, über die drei Testzeitpunkte mit „trifft eher nicht zu“ von den Schülern beantwortet. Bei der Frage, ob der Lernaufwand bei Multiple Choice Fragen niedrig ist, konnte eine Veränderung über die Zeit festgestellt werden. Die Schüler tendierten beim ersten Zeitpunkt dazu „trifft eher zu“ und beim dritten Zeitpunkt „trifft eher nicht zu“ anzukreuzen.

Im Schuljahr 2006/2007 wurden in der 2A Veränderungen über die Zeit bei den Fragen nach der Einfachheit der Fragen, des Stoffgebietes und dem Lernaufwand bei Multiple Choice Tests beobachtet. Bei der Frage ob die Fragen einfach waren, gaben die Schüler zu beiden Zeitpunkten „trifft eher zu“ an, wobei es zum ersten Zeitpunkt Abweichungen in Richtung „trifft eher nicht zu“ und zum dritten Zeitpunkt in Richtung „trifft zu“ gab. Bei der Einfachheit des Stoffgebietes tendierten die Schüler zum ersten Zeitpunkt zu „trifft eher nicht zu“ und beim dritten zu „trifft eher zu“. „Der Lernaufwand bei Multiple Choice Tests ist niedrig“ wurde beim

ersten Zeitpunkt mit „trifft eher nicht zu“ und beim dritten mit „trifft eher zu“ beantwortet. Der Lernaufwand bei offenen Fragen wurde über alle drei Zeitpunkte als nicht hoch eingestuft.

In der 2B gaben die Schüler bei der Einfachheit der Fragen zu beiden Zeitpunkten „trifft eher zu“ an, wobei es zum ersten Zeitpunkt Abweichungen in Richtung „trifft eher nicht zu“ und zum vierten Zeitpunkt in Richtung „trifft eher zu“ gab. Die Schüler beurteilten die Einfachheit des Stoffgebietes mit „trifft eher nicht zu“, wobei es zum ersten Zeitpunkt Abweichungen in Richtung „trifft eher nicht zu“ und zum vierten Zeitpunkt in Richtung „trifft eher zu“ gab. Ob der Lernaufwand bei Multiple Choice Tests niedrig ist, wurde über alle Zeitpunkte mit „trifft eher zu“ und ob er bei offenen Fragen hoch ist mit „trifft eher nicht zu“ beantwortet.

In der 4AF im Schuljahr 2006/2007 gab es keine Unterschiede im Antwortverhalten der Schüler über die vier Zeitpunkte. Sie schätzten die Fragen als eher einfach, das Stoffgebiet als eher einfach ein und tendierten dazu, dass der Lernaufwand bei Multiple Choice Tests eher niedrig und bei offenen Fragen eher nicht hoch ist.

Im Schuljahr 2005/2006 konnten in der 2B Unterschiede im Antwortverhalten über die Zeit beobachtet werden. Die Frage nach der Einfachheit der Fragen wurde zum ersten Zeitpunkt mit „trifft eher zu“ dagegen zum dritten mit „trifft eher nicht zu“ beantwortet. Das Stoffgebiet wurde über alle Zeitpunkte als nicht so einfach eingestuft, wobei es zum ersten Zeitpunkt Abweichungen in Richtung „trifft eher zu“ und zum dritten Zeitpunkt in Richtung „trifft eher nicht zu“ gab. Ob der Lernaufwand bei Multiple Choice Tests niedrig ist, wurde von den Schülern bei der ersten Testung eher bejaht, bei der letzten eher verneint. Den Lernaufwand bei offenen Fragen schätzten die Schüler zum ersten Zeitpunkt als eher nicht hoch bei letzten als nicht hoch ein.

In der 3A gaben die Schüler an, dass die Fragen über alle drei Zeitpunkte eher nicht einfach waren und dass der Lernaufwand für offene Fragen nicht hoch ist. Das Stoffgebiet schätzten sie beim ersten Testzeitpunkt als eher nicht einfach beim dritten als nicht einfach ein. Der Lernaufwand bei Multiple Choice Tests

wurde zum ersten Zeitpunkt als eher niedrig und beim dritten als eher nicht niedrig bewertet.

In der 5A empfanden die Schüler bei der ersten Testung die Fragen als eher einfach bei der vierten dagegen als eher nicht einfach. Die Schüler gaben zum ersten Zeitpunkt an, dass das Stoffgebiet eher einfach und beim vierten das es nicht einfach war. Der Lernaufwand bei Multiple Choice Tests wurde zum ersten Zeitpunkt als eher niedrig und beim vierten als eher nicht niedrig bewertet. Der Lernaufwand für offene Fragen wurde auch in dieser Klasse als nicht hoch eingeschätzt.

III.4.3 Veränderungen über die Zeit, bei zwei Testzeitpunkten

Im Schuljahr 2007/2008 konnten keine Veränderungen in der 2A über die Zeit bei den Fragen nach der Einfachheit der Fragen, des Stoffgebietes und dem Lernaufwand bei Multiple Choice Tests gefunden werden. Bei der Frage nach der Einfachheit der gestellten Fragen ergaben sich zwischen den zwei Zeitpunkten keine signifikanten Ergebnisse, wobei die Antworten zwischen „trifft eher zu“ und „trifft eher nicht zu“ lagen. Das Stoffgebiet wurde bei beiden Testungen als eher nicht so einfach und der Lernaufwand bei Multiple Choice Tests als eher nicht zu niedrig beurteilt. Der Lernaufwand bei offenen Fragen ist hoch, wurde zu beiden Zeitpunkten mit „trifft eher nicht zu“ beantwortet, wobei es zum ersten Zeitpunkt Abweichungen in Richtung „trifft eher nicht zu“ und zum zweiten Zeitpunkt in Richtung „trifft eher zu“ gab.

In der 2B gaben die Schüler bei der Einfachheit der Fragen zu beiden Zeitpunkten „trifft eher zu“ an, wobei es zum ersten Zeitpunkt Abweichungen in Richtung „trifft eher zu“ und zum zweiten Zeitpunkt in Richtung „trifft eher nicht zu“ gab. Ob der Lernaufwand bei Multiple Choice Tests niedrig ist, wurde von den Schülern bei beiden Testungen eher bejaht. In dieser Klasse wurden die Fragen und das Stoffgebiet als eher einfach und der Lernaufwand bei offenen Fragen als nicht so hoch eingeschätzt.

Im Schuljahr 2006/2007 wurden in der 5A die Fragen des Tests über die Zeit als eher einfach und das Stoffgebiet als eher nicht einfach eingeschätzt. In diesem Schuljahr wurde die Frage, ob der Lernaufwand bei offenen Fragen hoch ist, über die zwei Testzeitpunkte mit „trifft eher nicht zu“ von den Schülern beantwortet. Bei der Frage, ob der Lernaufwand bei Multiple Choice Fragen niedrig ist, konnte eine Veränderung über die Zeit festgestellt werden. Die Schüler tendierten beim ersten Zeitpunkt dazu „trifft eher zu“ und beim zweiten Zeitpunkt „trifft eher nicht zu“ anzukreuzen.

Im Schuljahr 2005/2006 veränderten sich die Antworten der Schüler der 2A, über die Zeit, bei der Frage nach der Einfachheit der Fragen und ob der Lernaufwand bei Multiple Choice Tests niedrig ist. Zum ersten Zeitpunkt wurden die Fragen als eher einfach und der Lernaufwand bei den Tests als eher niedrig wahrgenommen. Beim zweiten Test wurden die Fragen als nicht so einfach und der Lernaufwand nicht so niedrig eingeschätzt. Das Stoffgebiet wurde in dieser Klasse stets als nicht so einfach und der Lernaufwand bei offenen Fragen als nicht hoch beurteilt.

In der 4AF wurde die Frage nach der Einfachheit der Fragen zum ersten Zeitpunkt von den Schülern mit „trifft eher zu“ und zum zweiten Zeitpunkt mit „trifft eher nicht zu“ beantwortet. Bei der Frage, ob der Lernaufwand bei Multiple Choice Fragen niedrig ist, konnte eine Veränderung über die Zeit festgestellt werden. Die Schüler tendierten beim ersten Zeitpunkt dazu „trifft eher zu“ und beim zweiten Zeitpunkt „trifft eher nicht zu“ anzukreuzen. Bei der Frage nach der Einfachheit des Stoffgebietes ergaben sich zwischen den zwei Zeitpunkten keine signifikanten Ergebnisse, wobei die Antworten zwischen „trifft eher zu“ und „trifft eher nicht zu“ lagen. Die Frage, ob der Lernaufwand bei offenen Fragen hoch ist, wurde über die zwei Testzeitpunkte mit „trifft nicht zu“ von den Schülern beantwortet.

III.5 Interpretation

Alle gefundenen Ergebnisse sind Rückmeldungen der befragten Schüler auf den jeweils vorangegangenen Multiple Choice Test. Wie bereits erwähnt können keine Verallgemeinerungen gemacht werden, da sich alle Daten und Ergebnisse auf die erhobene Population beziehen.

Wie wird der vorgegebene Multiple Choice Test von den Schülern eingeschätzt?

Es konnte gezeigt werden, dass gute Schüler die Fragen des vorgegebenen Multiple Choice Tests sowie das geprüfte Stoffgebiet als einfach wahrnahmen. Dies könnte darauf zurückzuführen sein, dass erfolgsmotivierte Schüler der Ansicht sind, sie seien gut in einem Fach, sie haben viel gelernt oder die Fragen waren wirklich einfach gestellt (GAGE, BERLINER 1996, 348ff, WEINER 1984, 258ff). Dies entspricht auch den Untersuchungen von GAGE und BERLINER (1996, 348ff), die beobachtet haben, dass die Schüler die eigene Anstrengung und den Schwierigkeitsgrad der Aufgaben für ihren Erfolg verantwortlich machen.

Bei welchem Prüfungsformat geben die Schüler an mehr lernen zu müssen?

Gute Schüler schätzen den Lernaufwand für Multiple Choice Tests niedriger ein als bei offenen Fragen. Bei einer Untersuchung von GANNESHOFFER und KIFFEL (2005, 62ff) wurden Hinweise gefunden, dass sich Studenten auf Multiple Choice Prüfungen primär mit Musterklausuren und mit Hilfe des „Wiedererkennfaktors“ vorbereiten. Bei offenen Fragen wird jedoch laut den Befragten auf Verständnis gelernt und mit Hilfe von Zusammenfassungen. Das heißt, dass der Lernaufwand ein größerer ist und daher auch eventuell mehr gelernt wird.

Dieses Ergebnis widerspricht jedoch den Ansichten von SCHINDLER (1990, 24), der der Meinung ist, dass der Lernaufwand bei Multiple Choice Fragen nicht unterschätzt werden soll, da man eine klare Entscheidung treffen muss, um an eine Antwort zu gelangen. "By causing students to worry about erroneous factors

that they otherwise would not have taken into consideration". (CHAN, KENNEDY 2002, 970)

Welche Form des Prüfens schätzen die befragten Schüler als schwieriger ein?

Schüler, die eine hohe Punkteanzahl bei dem Test erzielten, beurteilten offene Fragen als schwieriger und sind auch der Ansicht, dass das Wissen bei offenen Fragen länger hält. Bei der Frage nach der Dauerhaftigkeit des Prüfungswissens konnten in einer Studie beobachtet werden, dass es keine Unterschiede zwischen den beiden Formaten gibt (GANNESHOFFER, KIFFEL 2005, 62ff), was wiederum den Ergebnissen in meiner Arbeit widerspricht. Dies ist darauf zurückzuführen, dass sich Schüler von persönlichen Vorlieben, Abneigungen, Erwartungshaltungen, Wunschvorstellungen, sowie Entschuldigungsstrategien für schlechte Leistungen in ihrem Urteil leiten lassen und dieses somit verzerrt wird (STOLZ 1997, 124).

Das Ergebnis, dass Schüler mit guten Testleistungen, offene Fragen als schwieriger empfinden, könnte daran liegen, dass offene Aufgaben gekennzeichnet sind durch freie, nach eigenem Ermessen gegebene Antworten und das bei diesem Format Schüler bevorzugt werden, die sich schriftlich besser ausdrücken können (EBBINGHAUS, SCHMIDT 1999, 70ff).

Welches Prüfungsformat bevorzugen die befragten Schüler?

Bei einer Untersuchung von SCOLLER (1998, 468f) konnte beobachtet werden, dass die Präferenz für ein Format einen Einfluss auf das abschneiden bei diesem hat, was wiederum in meiner Arbeit nicht bestätigt werden konnte, denn es wurde nicht festgestellt, ob Schüler, die beim Multiple Choice Test gut abschneiden, ein Prüfungsformat bevorzugen. Eine Befragung von Psychologiestudenten brachte zum Vorschein, dass sie Kurzantwortfragen, gefolgt von Multiple Choice Fragen bevorzugen (NIELD, WINTRE 2002, 95).

Es zeigte sich an anderer Stelle in der Literatur, dass die Befragten Multiple Choice Fragen favorisierten, da sie bei diesen gegebenenfalls raten konnten und ihrer Meinung nach wenig geistige und körperliche Anstrengung erforderten

(NIELD, WINTRE 2002, 96ff). Präferenz für Kurzantworten zeigte sich dann, wenn Studenten mit diesen zeigen konnten, was sie gelernt hatten (NIELD, WINTRE 2002, 96ff).

Explorativ stellte ich mir noch die Frage, **inwiefern sich das Antwortverhalten von Schülern über den Zeitraum von einem Schuljahr veränderte**. Aufgrund der Daten konnten wie bereits erwähnt nur vier Feedbackfragen für diese Fragestellung herangezogen werden. Erstens die Frage nach der Einfachheit der Fragen, zweitens nach der Einfachheit des Stoffgebietes, drittens die Frage, ob der Lernaufwand für Multiple Choice Fragen niedrig ist und viertens, ob der Lernaufwand für offene Fragen hoch ist.

Die Frage nach der Einfachheit änderte sich ab und an über die Testzeitpunkte in einem Schuljahr, was daran liegen könnte, dass zu jedem Zeitpunkt ein anderer Test vorgegeben wurde. Die Antworten variierten von „trifft eher zu“ bis „trifft eher nicht zu“. Dieselben Veränderungen konnten teilweise bei der Frage nach der Einfachheit des Stoffgebietes beobachtet werden, wobei dieses sich auch über ein Schuljahr hindurch veränderte. Über alle Klassen veränderten die Schüler ihre Antworten auf die Frage, ob der Lernaufwand bei Multiple Choice Tests niedrig ist, insgesamt sieben Mal signifikant von „trifft eher zu“ auf „trifft eher nicht zu“ oder umgekehrt. Auf der einen Seite waren die geprüften Multiple Choice Tests zu den verschiedenen Zeitpunkten nicht gleich, auf der anderen Seite konnten bereits in der Literatur Hinweise darauf gefunden werden, dass die verschiedenen Aufgabenformen im Unterricht geübt werden müssen, da sonst die Bewältigung der ungewohnten Aufgabenform den Schülern Schwierigkeiten bereitet (SACHER 2004, 68). Aufgrund des Trainings mit Multiple Choice Items waren die Leistungen beim Multiple Choice Test höher, als beim Training mit “recall” Items, wie sich in einer Studie von SAX und LEVERNE (1968, 172) zeigte.

Bei meinem Datensatz könnte ebenfalls ein Trainingseffekt über die Zeit aufgetreten sein. Auch FICHTEN (1993, 98f) konnte Tendenzen finden, dass erst methodische Kompetenz aufgebaut werden muss, um positive Erfahrungen bei den Schülern zu erzeugen und somit deren Bewertungen langfristig und nachhaltig zu beeinflussen. Prüfungsroutine verhilft den Schülern zu einer

effektiveren Meisterung der Aufgaben und sie fühlen sich in der Prüfungssituation sicherer (SALZGER 1998, 82). In Untersuchungen über Prüfungserfahrung konnte ein positiver Zusammenhang zwischen Testerfahrung und Testleistung gezeigt werden. Bereits bei der zweimaligen Bearbeitung ein und desselben Tests konnten Leistungssteigerungen beobachtet werden, die auf die Übung mit der Aufgabenart und der Testsituation zurück zu führen sind (SCHNEIDER 1987, 5f).

Schüler, die gebundene Aufgaben bereits kennen, haben den Vorteil zu wissen, wie an die Aufgaben herangegangen wird. „So beherrschen sie die Technik, wie man Informationen einsetzt, um falsche Antworten zu eliminieren und die richtigen zu deduzieren.“ (SCHINDLER 1990, 52).

Über alle Klassen veränderte sich die Antwort auf die Frage, ob der Lernaufwand bei offenen Fragen hoch ist nur zweimal, von „trifft eher nicht zu“ auf „trifft nicht zu“. Die Schüler schätzten nach einem Multiple Choice Test, den Lernaufwand für offene Fragen nicht hoch ein. Dies könnte daran liegen, dass es kein Bezugsereignis, offene Fragen, gab und die Antworten Rückmeldungen auf den vorangegangenen Test darstellten.

III.6 Diskussion

Zu bedenken ist, dass man sich mit wachsender Aufgabenzahl der Ermüdungsgrenze nähert und dann nicht mehr ausschließlich Fachleistung sondern auch allgemeine Belastbarkeit misst (SACHER 2004, 122f).

Bei einer Studie von KRENN (1994, 118ff) konnte beobachtet werden, dass Schüler, die schlecht bei einem Konzentrationsleistungstest abschneiden, auch in anderen Lebensbereichen Konzentrationsfehler aufweisen und Schüler, die schlechter motiviert sind, in Schularbeiten schlechter abschneiden. Weiters konnte er herausfinden, dass Schüler der Ansicht sind, ihre Leistungen seien abhängig von der Beziehung zum Lehrer. Konkreter gesagt, schlechte Schüler vermuten weniger gemocht zu werden. MOOS und BABAD (1979, 1996; zit. nach CLAUSEN

2002, 44) konnten feststellen, dass Schüler mit guten Noten ihre Umwelt, beziehungsweise den Unterricht, positiver beurteilen als schlechte Schüler.

Bei all den Ergebnissen ist die Problematik der Antwortstile nicht ganz außer acht zu lassen. Die Ja- Tendenz bezeichnet die Neigung auf Fragen zustimmende Antworten zu geben (DREESMANN 1982, 110), die Tendenz zur Mitte wurde in meiner Arbeit versucht auszuschalten, indem eine gerade Zahl von Antworten gegeben wurde. Trotzdem musste ich feststellen, dass die Schüler eher zu den Antworten in der Mitte tendierten. Auch ATTALI und BAR-HILLEL (2003, 111ff) konnten bei diversen Versuchen beobachten, dass die Befragten bei ihren Antworten zur Mitte neigten. Bei der geschlossenen Aufgabenform besteht ein Nachteil darin, dass der Befragte das Gefühl hat, die Antworten bringen nicht das zum Ausdruck was sie eigentlich wollen und müssen daher gewisse Kompromisse eingehen (MITTELSTÄDT 2006, 78). Andere Untersuchungen zeigen, dass Studenten hilflos und frustriert sind, wenn es darum geht, eine Antwort, und zwar die beste, anzukreuzen (NIELD, WINTRE 2002, 95). Dies könnte sich auch auf die Antworten beim Feedbackbogen auswirken, was jedoch in meiner Arbeit nicht untersucht wurde.

Zuletzt soll noch die soziale Erwünschtheit genannt werden, bei der die Schüler dazu neigen, so zu antworten, wie es von ihnen erwartet wird (DREESMANN 1982, 110). Dies könnte auch in meiner Arbeit der Fall sein, da die Schüler den Feedbackbogen nach dem Test ausfüllten und abgaben. Vielleicht wollten sie den Lehrer nicht „verärgern“.

III.7 Resümee und Ausblick

Wie sich im theoretischen Teil meiner Arbeit zeigte, findet das Thema Multiple Choice Test im deutschsprachigen Raum kaum Beachtung. Es werden zwar Fragen bei PISA in diesem Prüfungsformat gestellt, in der Schulpraxis findet diese Art der Aufgaben jedoch kaum Beachtung. Die meisten Studien zum Thema Multiple Choice Tests und zu Tests mit offenen Fragen, deuten darauf hin, dass

Aufgabenformate dem Befragten leichter fallen, wenn sie vorher geübt wurden. Die meisten Studien deuten des Weiteren darauf hin, dass es sowohl Schüler und Studenten gibt die ein Format bevorzugen. Außerdem sollte man nicht vernachlässigen, dass die Wahl des Antwortformats vom Prüfer selbstständig festgelegt wird.

Ziel meiner Untersuchung war es, sich näher mit der Sicht der Schüler auseinander zu setzen, um einen Eindruck davon zu gewinnen, wie diese Multiple Choice Tests einschätzen. Mir ging es also um die Sicht der Experten, die mit diesem Format geprüft und deren Leistungen mittels diesem festgestellt werden. Sicherlich stellen diese Art der Prüfung andere Anforderungen an die Befragten als herkömmliche Prüfungen, denn bei offenen Fragen, lag es in der Hand des Schülers, was er als Antwort gibt. In den einzelnen Schritten die ich untersucht habe, konnten Tendenzen festgestellt werden, die über weite Strecken sehr ähnlich sind aber nicht über alle Schüler einheitlich. Es wurden zwar viele Schüler befragt und das über teilweise drei Jahre, es handelte sich jedoch immer um kleine Klassen, in unterschiedlichsten Fächern.

Die Ergebnisse dieser Arbeit deuten darauf hin, dass Schüler die bei einem Multiple Choice Test gut abschneiden, diesen als leichter empfinden, sie der Ansicht sind, sie müssten für Tests mit offenen Fragen mehr lernen und dass sie sich bei dieser Form länger an das Gelernte erinnern können, sie jedoch kein Aufgabenform bevorzugen. In vieler Hinsicht widersprechen diese Ergebnisse der gelesenen Literatur, was daran liegen könnte, dass viele Studien zu diesem Thema aus dem amerikanischen Raum stammen, wo die Prüfungspraxis in der Schule eine andere ist als in Österreich und die Schüler sehr früh mit Multiple Choice Tests konfrontiert werden.

Die Meinungen der Schüler änderten sich im Verlauf eines Schuljahres kaum und wenn doch dann von „trifft eher zu“ auf „trifft eher nicht zu“. Dies deutet auf eine gewisse Tendenz in der Grundhaltung hin; entweder der Schüler findet von Anfang an Multiple Choice Tests leichter oder eben nicht.

Sicherlich bedarf es auf diesem Gebiet noch einiges an Untersuchungen, trotzdem gibt es Hinweise darauf, dass der Umgang mit Prüfungsfragen

Übungssache ist. Daher ist es nicht verwunderlich, dass sich Schüler bei diesem Format anfänglich etwas schwerer tun und Studenten bei diversen Eingangstests allein vom Format her unter Umständen irritiert sind. In zukünftigen Untersuchungen gilt es also zu klären wie sinnvoll es ist, ein ungeübtes Aufgabenformat wie Multiple Choice Fragen, in einem internationalen Leistungsvergleich zu verwenden und sich dann zu erwarten, dass die Ergebnisse unabhängig von der Art der Frage seien.

Appendix A

Feedbackfragen

„1) Zeitfaktor: Die Zeit für den Test war 'ausreichend'?

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

2) Die Fragen waren 'einfach'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

3) Die Formulierung der Fragen war 'verständlich' und 'klar'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

4) Die Formulierung der Antworten war 'verständlich' und 'klar'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

5) Der Stoffgebietsfortschritt bzw. der Vortrag im Unterricht war 'zu schnell'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

6) Die Aufbereitung des Stoffgebietes bzw. der Vortrag im Unterricht war 'gut'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

7) Das Stoffgebiet ist 'einfach'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

8) Das Stoffgebiet ist 'interessant'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

9) Der Nutzen des Stoffgebietes ist für eine spätere Berufstätigkeit 'erkennbar'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

10) Die Menge der Fallbeispiele (Übungsbeispiele, Übungsprogramme) war 'ausreichend'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

11) Der Praxisbezug, d.h. die Verweise zur Praxis (z.B. in der Praxis wird das so ... gemacht) war 'ausreichend'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

12) Der Lernaufwand bei 'Multiple Choice Tests' ist 'niedrig'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

13) Der Lernaufwand bei 'offenen Fragen' (z.B. Matheschularbeit; keine vorgefertigten Antworten) ist 'hoch'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

14) Vergessenskurve: Wann ist das Wissen für diese Prüfung 'weg'?

1-Tag

3-Tage

1-Woche

1-Monat

1/2-Jahr

1-Jahr oder mehr

15) Der Lernaufwand für diese Prüfung war insgesamt (inkl. Vorbereitung für Wiederholungen)

15-Minuten

1/2-Stunde

1-Stunde

3-Stunden

5-Stunden

10-Stunden oder mehr

16) Der Schwierigkeitsgrad der Tests ist 'höher' bei

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

17) Der Lernaufwand ist 'höher' bei

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

18) Vergessenskurve: Wobei 'hält' das Wissen 'länger'?

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

19) Welche Tests sind Ihnen 'lieber'?

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

20) Kommunikationsmöglichkeiten (Schummeln) sind gegeben durch folgende Mittel

Email

Sonstige Programme unter Windows

Handzettel

Kommunikation mit dem Nachbar

Andere

Keine

21) Kommunikationsmöglichkeiten (Schummeln) sind 'einfach'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu

22) 'Bessere' Schummelmöglichkeit ist möglich bei

Multiple Choice Test

Offene Fragen (z.B. Matheschularbeit; keine vorgefertigten Antworten)

23) Referate sind 'erwünscht'

Trifft zu

Trifft eher zu

Trifft eher nicht zu

Trifft nicht zu“ (FÜHRER 2007, 258f)

Appendix B

Schuljahr	Klasse	Fach	Anzahl der Tests	Anzahl der Schüler
2007/2008	2A	AINF	2	27
2007/2008	2B	AINF	2	20
2007/2008	3A	INIT	5	26
2007/2008	4A	PRE	3	22
2007/2008	4AF	BUN	3	18
2007/2008	5A	INIT	3	21
2006/2007	2A	AINF	3	25
2006/2007	2B	AINF	4	32
2006/2007	3A	INIT	4	25
2006/2007	3A	PRE	3	25
2006/2007	4AF	BUN	4	23
2006/2007	5A	INIT	2	25
2005/2006	2A	AINF	2	27
2005/2006	2B	AINF	3	28
2005/2006	3A	INIT	3	24
2005/2006	4AF	BUN	2	20
2005/2006	5A	INIT	2	26
2005/2006	5A	PRE	2	26

Tabelle 17: Testplan über alle 3 Schuljahre

Literaturverzeichnis

- ATTALI, Y., BAR-HILLEL, M. (2003): Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. In: Journal of educational measurement, Vol. 40, No. 2, Summer, 109-128.
- BALCH, W. R. (2002): Item order affects performance on multiple-choice exams. In: GRIGGS, R. A. (Hrsg.) (2002): Handbook for teaching introductory psychology. With an emphasis on assessment. Volume III. New Jersey: Lawrence Erlbaum Associates, Inc., 106-108.
- BAR-HILLEL, M., BUDESCU, D., ATTALI, Y. (2005): Scoring and keying multiple choice tests: a case study irrationality. In: Mind & Society, Vol. 4, 3-12.
- BAUMERT, J., LEHMANN, R., u.a. (1997): TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske + Budrich.
- BESSOTH, R. (2002): Lernerfolge – in der Zukunft eine zentrale Verantwortung deutscher Schulleitungen. Lerneinheit 45.01. In: BESSOTH, R. (Hrsg.) (1989/2002 aktualisiert): Schulleitung. Ein Lernsystem. Pädagogische Führung. Unterricht 4a. Neuwied: Hermann Luchterhand Verlag.
- BIRENBAUM, M., TATSUOKA, K. K. (1987): Open-ended versus multiple-choice response formats – it does make a difference for diagnostic purposes. In: Applied psychological measurement, Vol. 11, No. 4, December, 385-395.
- BORENSTEIN, M., HEDGES, L., u.a. (2005): Comprehensive Meta-analysis Version 2. Englewood NJ: Biostat. Im Internet unter www.meta-analysis.com [05.05.2009].
- BOSSONG, B. (2006): Wie sympathisch bin ich meiner Lehrerin und wie schätzt sie meine Fähigkeiten ein? Zum Einfluss wahrgenommener Lehrereinstellungen auf Selbstkonzepte der Fähigkeit bei Grundschulern. In: HOSENFELD, I., SCHRADER, F.-W. (Hrsg.) (2006): Schulische Leistung. Grundlagen, Bedingungen, Perspektiven. Münster: Waxmann Verlag GmbH, 263-274.
- BRIDGEMANN, B. (1992): A comparison of quantitative questions in open-ended and multiple-choice format. In: Journal of educational measurement, Vol. 29, No. 3, Fall, 253-271.
- BUSE, L. (1977): Die Abhängigkeit des Reliabilitätskoeffizienten von Rateinflüssen. In: Zeitschrift für experimentelle und angewandte Psychologie, Band 24, Heft 4, 546-558.

- BUTLER, A. C., KARPICKE, J. D., ROEDIGER, H. L. III (2007): The effect of type and timing of feedback on learning from multiple-choice tests. In: Journal of Experimental Psychology: Applied, Vol. 13, No. 4, 273-281.
- CHAN, N., KENNEDY, P. E. (2002): Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. In: Southern Economic Journal, Vol. 68, No. 4, 957-971.
- CLAUSEN, M. (2002): Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität. Münster: Waxmann Verlag GmbH. (Pädagogische Psychologie und Entwicklungspsychologie, Band 29).
- CZERWENKA, K., NÖLLE, K., u.a. (1990): Schülerurteile über die Schule. Bericht über eine internationale Untersuchung. Frankfurt am Main; Wien [u.a.]: Peter Lang Verlag. (Europäische Hochschulschriften: Reihe 11, Pädagogik, Band 419).
- DITTON, H. (2002): Lehrkräfte und Unterricht aus Schülersicht. Ergebnisse einer Untersuchung im Fach Mathematik. In: Zeitschrift für Pädagogik, 48. Jahrgang, Heft 2, 262-286.
- DREESMANN, H. (1982): Unterrichtsklima. Wie Schüler den Unterricht wahrnehmen. Ein Beitrag zur „Ökologie des Lernens“. Weinheim [u.a.]: Beltz Verlag. (U-&-S-Pädagogik: Studentexte Päd/Psy).
- DUFRESNE, R. J., LEONARD, W. J., GERACE, W. J. (2002): Making sense of students' answers to multiple-choice questions. In: The Physics Teacher, Vol. 40, March, 174-180.
- EBBINGHAUS, M., SCHMIDT, J. U. (1999): Prüfungsmethoden und Aufgabenarten. Hrsg.: Bundesinstitut für Berufsbildung, Der Generalsekretär. Bielefeld: Bertelsmann.
- FICHTEN, W. (1993): Unterricht aus Schülersicht. Die Schülerwahrnehmung von Unterricht als erziehungswissenschaftlicher Gegenstand und ihre Verarbeitung im Unterricht. Frankfurt am Main; Wien [u.a.]: Peter Lang Verlag. (Europäische Hochschulschriften: Reihe 11, Pädagogik, Band 562).
- FRIEDRICH, H. F. (1987): Multiple Choice- Prüfungen im Lehr- Lernsystem „Funkkolleg“. In: HORN, R., INGENKAMP, K., JÄGER, R. S. (1987): Tests und Trends 6. Jahrbuch der Pädagogischen Diagnostik. München und Weinheim: Psychologie Verlags Union, 128-175.
- FRIEDRICHS, J. (1980): Methoden empirischer Sozialforschung. Opladen: Westdeutscher Verlag GmbH.

- FÜHRER, W. (2007): Auswirkungen verschiedener zeitlicher Ausgangssituationen auf das Testergebnis eines Multiple Choice Tests einschließlich einer didaktischen Bewertung und Reflexion der Konsequenzen für einen zeitgemäßen Mathematikunterricht. Unveröffentlichte Dissertation Wien: Universität.
- GAGE, N. L., BERLINER, D. C. (1986): Pädagogische Psychologie. 4. völlig neu bearbeitete Auflage. Weinheim: Beltz, Psychologie Verlags Union.
- GAGE, N. L., BERLINER, D. C. (1996): Pädagogische Psychologie. 5. vollständig überarbeitete Auflage. Weinheim: Beltz, Psychologie Verlags Union.
- GANNESHOFFER, K., KIFFEL, A. (2005): Einschätzung der Multiple-Choice-Prüfungen aus der Sicht von Studierenden. Eine hochschuldidaktisch-qualitative Studie. Unveröffentlichte Diplomarbeit Wien: Wirtschaftsuniversität.
- GERSTENMAIER, J. (1975): Urteile von Schülern über Lehrer. Eine Analyse ausgewählter empirischer Untersuchungen. Weinheim [u.a.]: Beltz Verlag. (Erziehungswissenschaftliche Forschung, Band 6).
- GOETHALS, R. (1994): Die praktische Erprobung von Alternativen zur Multiple-Choice-Vorgabe bei Computertests. Erfahrungen mit der sequentiellen Distraktorenvorgabe und der kumulativen Lösungsmethode am Personal Computer. Unveröffentlichte Dissertation Wien: Universität.
- HAIDER, G., SCHREINER, C. (2006): PISA-Leistung und Schulnoten. In: HAIDER, G., SCHREINER, C. (Hrsg.) (2006): Die PISA - Studie. Österreichs Schulsystem im internationalen Wettbewerb. Wien [u.a.]: Böhlau Verlag, 229-236.
- HANISCH, G. (2003): Leistungsbeurteilung. In: BÁBOSIK, I., OLECHOWSKI, R. (Hrsg.) (2003): Lehren-Lernen-Prüfen. Frankfurt am Main; Wien [u.a.]: Peter Lang Verlag, 205-214. (Schule – Wissenschaft – Politik, Band 16).
- HAYNIE, W. J. III (1994): Effects of multiple-choice and short-answer tests on delayed retention learning. In: Journal of technology education, Vol. 6, No. 1, Fall, 32-44.
- HELLER, K. A., HANY, E. A. (2001): Standardisierte Schulleistungsmessungen. In: WEINERT, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. Weinheim [u.a.]: Beltz Verlag, 87-101.
- HEPP, G., BENZ, T. (1987): Möglichkeiten der Erfolgskontrolle im Unterricht. Beispiele für Schüleraktivierung jenseits von Abfragen und Klassenarbeit. In: Westermanns Pädagogische Beiträge, 39. Jahrgang, Heft 5, 29-35.
- HOSENFELD, I., SCHRADER, F.-W. (Hrsg.) (2006): Schulische Leistung. Grundlagen, Bedingungen, Perspektiven. Münster: Waxmann Verlag GmbH, 9-11.

- INGENKAMP, K. (1971): Die Entwicklung von Schultests. In: INGENKAMP, K. (Hrsg.) (1971): Tests in der Schulpraxis. Eine Einführung in Aufgabenstellung, Beurteilung und Anwendung von Tests. Weinheim [u.a.]: Beltz Verlag, 43-73.
- INGENKAMP, K. (1989): Diagnostik in der Schule. Beiträge zu Schlüsselfragen der Schülerbeurteilung. Weinheim und Basel: Beltz Verlag.
- JACHMANN, M., TILLMANN, K.-J. (2000): Sind Noten gerechter als Berichtszeugnisse? Wie Schüler, Lehrer und Eltern die schulische Beurteilungspraxis sehen. In: Pädagogik, 52. Jahrgang, Heft 9, 36-43.
- JÜRGENS, E. (2005): Leistung und Beurteilung in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht. 6. aktualisierte und stark erweiterte Auflage. Sankt Augustin: Academia Verlag.
- KLAUER, K. J. (2001): Wie misst man Schulleistungen? In: WEINERT, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. Weinheim [u.a.]: Beltz Verlag, 102-115.
- KLINGLER, J. (1982): Objektivität und Gerechtigkeit bei der Beurteilung von Schülerleistungen. Klagenfurt: Universitätsverlag Carinthia.
- KRENN, B. (1994): Bedeutung und Auswirkung der Leistungsbeurteilung auf die tatsächliche Leistung des Schülers. Eine empirische Untersuchung durchgeführt in Oberösterreich. Unveröffentlichte Diplomarbeit Wien: Universität.
- LBVO (BGBl. Nr. 371/1974, zuletzt geändert durch BGBl. II Nr. 35/1997) im Internet unter www.bmukk.gv.at [30.4.2009].
- LIEDTKE, M. (1991): Ist das Zeugnis das Armutszeugnis der Schule? In: HOHENZOLLERN, Prinz von, J. G., LIEDTKE, M. (Hrsg.) (1991): Schülerbeurteilungen und Schulzeugnisse. Historische und systematische Aspekte. Heilbrunn/Obb.: Verlag Julius Klinkhardt, 25-36. (Schriftenreihe zum Bayerischen Schulmuseum Ichenhausen, Band 10).
- LIENERT, G. A., RAATZ, U. (1998): Testaufbau und Testanalyse. 6. Auflage. Weinheim: Beltz Psychologie Verlags Union.
- LISSMANN, U. (1998): Probleme und Möglichkeiten der Schülerbeurteilung – Folienatlas. Landau: Verlag Empirische Pädagogik. (Materialien für Lehre, Aus- und Weiterbildung, Band 17).
- MEMMERT, W. (1997): Kunstfehler beim Unterrichten. In: SCHWARZ, B., PRANGE, K. (Hrsg.) (1997): Schlechte Lehrer/innen. Zu einem vernachlässigten Aspekt des Lehrberufs. Weinheim und Basel: Beltz Verlag, 248-274.

- MEYERHÖFER, W. (2003): Was testen Tests? Objektiv-hermeneutische Analyse am Beispiel von TIMSS und PISA. Unveröffentlichte Dissertation Potsdam: Universität.
- MITTELSTÄDT, H. (2006): Evaluation von Unterricht und Schule. Strategien und Praxistipps. Mülheim an der Ruhr: Verlag an der Ruhr.
- NEELY, D. L., SPRINGSTON, F. J., MCCANN, S. J. H. (2002): Does item order affect performance on multiple-choice exams? In: GRIGGS, R. A. (Hrsg.) (2002): Handbook for teaching introductory psychology. With an emphasis on assessment. Volume III. New Jersey: Lawrence Erlbaum Associates, Inc., 108-110.
- NIELD, A. F., WINTRE, M. G. (2002): Multiple-Choice questions with an option to comment: Student attitudes and use. In: GRIGGS, R. A. (Hrsg.) (2002): Handbook for teaching introductory psychology. With an emphasis on assessment. Volume III. New Jersey: Lawrence Erlbaum Associates, Inc., 95-99.
- OWENS, R. E., HANNA, G. S., COPPEDGE, F. L. (1970): Comparison of multiple-choice tests using different types of distractor selection techniques. In: Journal of educational measurement, Vol. 7, No. 2, Summer, 87-90.
- PERRIN, G. (2004): Multiple choice – making the best of it. In: Fremdsprachenunterricht, Band 1, Heft 3, 171-175.
- PESCHEK, W., SCHNEIDER, E. (2006): PISA Mathematik: Die österreichischen Ergebnisse aus fachdidaktischer Sicht. In: HAIDER, G., SCHREINER, C. (Hrsg.) (2006): Die PISA - Studie. Österreichs Schulsystem im internationalen Wettbewerb. Wien [u.a.]: Böhlau Verlag, 73-84.
- PLATH, I. (2006): Metaanalyse. In: ROST, D. H. (Hrsg.) (2006): Handwörterbuch Pädagogische Psychologie. 3. überarbeitete und erweiterte Auflage. Weinheim und Basel: Beltz Verlag, 474-480.
- PÖGGELER, F. (1991): Zur Relation von Schulerfolg und Lebenserfolg – Historische und aktuelle Aspekte. In: HOENZOLLERN, Prinz von, J. G., LIEDTKE, M. (Hrsg.) (1991): Schülerbeurteilungen und Schulzeugnisse. Historische und systematische Aspekte. Heilbrunn/Obb.: Verlag Julius Klinkhardt, 209-224. (Schriftenreihe zum Bayerischen Schulmuseum Ichenhausen, Band 10).
- RAPP, G. (1971): Was prüft ein Leistungstest? In: INGENKAMP, K. (Hrsg.) (1971): Tests in der Schulpraxis. Eine Einführung in Aufgabenstellung, Beurteilung und Anwendung von Tests. Weinheim [u.a.]: Beltz Verlag, 74-78.

- ROLLETT, B. (1997): Lernen und Lehren. Eine Einführung in die Pädagogische Psychologie und ihre entwicklungspsychologischen Grundlagen. 5. überarbeitete Auflage. Wien: WUV-Universitätsverlag. (WUV Studienbücher Psychologie, Band 2).
- ROWLEY, G. L. (1974): Which examinees are most favoured by the use of multiple choice tests? In: Journal of educational measurement, Vol. 11, No. 1, Spring, 15-23.
- SACHER, W. (2004): Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe. 4. überarbeitete und erweiterte Auflage. Bad Heilbrunn/Obb.: Verlag Julius Klinkhardt.
- SALZGER, B. (1998): Mathematik und Deutsch aus der Sicht von Schülerinnen und Schülern. Eine empirische Untersuchung in der 7. und 8. Schulstufe. Unveröffentlichte Dissertation Wien: Universität.
- SAX, G., LEVERNE, S. C. (1968): An empirical comparison of the effects of recall and multiple-choice tests on student achievement. In: Journal of educational measurement, Vol. 5, No. 3, Summer, 169-173.
- SCANNELL, D. P., TRACY, D. B. (1977): Testen und Messen im Unterricht. Wege zu einer differenzierten Überprüfung kognitiven und affektiven Lernens. Aus dem Amerikanischen übersetzt von Helmut Dreesmann und Astrid Schick. 1. Auflage. Weinheim und Basel: Beltz Verlag.
- SCHEIBER, B. (1990): Leistungsfeststellung und Leistungsbeurteilung in der Schule. In: OLECHOWSKI, R., RIEDER, K. (Hrsg.) (1990): Motivieren ohne Noten. Wien [u.a.]: Jugend und Volk, 92-128. (Schule – Wissenschaft – Politik, Band 3).
- SCHINDLER, M. (1990): Der Zusammenhang zwischen Prüfungsform und Lernverhalten. Unveröffentlichte Diplomarbeit Wien: Wirtschaftsuniversität.
- SCHNEIDER, B. (1987): Vorbereitung auf Intelligenz- und Leistungstests: Eine Gefahr für die Eignungsdiagnostik? In: HORN, R., INGENKAMP, K., JÄGER, R. S. (1987): Tests und Trends 6. Jahrbuch der Pädagogischen Diagnostik. München und Weinheim: Psychologie Verlags Union, 3-25.
- SchUG (1986) im Internet unter www.bmukk.gv.at [30.4.2009].
- SCHWENDENWEIN, W. (1985): Ausgewählte Prüfungsaspekte als Beitrag zum Problem der Leistungsbeurteilung. In: Erziehung und Unterricht, 135. Jahrgang, Heft 1-10, 149-157.
- SCHWENDENWEIN, W. (2000): Theorie des Unterrichtens und Prüfens. 7. Auflage. Wien: WUV-Universitätsverlag. (WUV Studienbücher Sozialwissenschaften, Band 4).

- SCOULLER, K. (1998): The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. In: Higher Education, Vol. 35, No. 4, 453-472.
- SCOULLER, K. M., PROSSER, M. (1994): Students' experiences in studying for multiple choice question examinations. In: Studies in Higher Education, Vol. 19, No. 3, 267-279.
- SMITH, M. III, WHITE, K. P., COOP, R. H. (1979): The effect of item type on the consequences of changing answers on multiple choice tests. In: Journal of educational measurement, Vol. 16, No. 3, Fall, 203-208.
- STOLZ, G. E. (1997): Der schlechte Lehrer aus der Sicht von Schülern. In: SCHWARZ, B., PRANGE, K. (Hrsg.) (1997): Schlechte Lehrer/innen. Zu einem vernachlässigten Aspekt des Lehrberufs. Weinheim und Basel: Beltz Verlag, 124-178.
- WEINER, B. (1984): Motivationspsychologie. Aus dem Amerikanischen übersetzt von Rainer Reisenzein unter Mitarbeit von Wilfried Prantner. Weinheim und Basel: Beltz Verlag.

Lebenslauf

Vorname: Claudia

Nachname: Hladecek

Geburtsdatum: 25.04.1981

Ausbildungsweg:

1987 bis 1991 Volksschule, Wien 1100

1992 bis 1999 AHS, Realgymnasium, Wien 1100, Matura

1999 bis 2003 Studium der Psychologie, Universität Wien

2003 bis 2009 Studium der Pädagogik, Universität Wien

Schwerpunkte: Schulpädagogik und Sonder- Heilpädagogik

Berufliche Erfahrungen:

2000 bis 2005 Publikumsdienst am „Wiener Burgtheater“

2002 bis 2007 Betreuerin von Kindern aus sozial benachteiligten Familien zwischen 6 und 14 Jahren bei der „Wiener Jugenderholung“
MA 11

2006/2007 Pädagogische Leitung bei der „Wiener Jugenderholung“

seit 2007 Betreuerin in einer Wohngemeinschaft beim Verein für
Gemeinwesenintegration und Normalisierung

Praxis:

- Wissenschaftliches Praktikum am Ludwig Boltzmann- Institut für Schulentwicklung und international vergleichende Schulforschung
- Praxis im pädagogischen Bereich beim Verein Wiener Jugenderholung als Betreuerin und pädagogische Leitung